DOKTORANDSKÉ DNY 2019

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

15. a 22. listopadu 2019

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2019
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

Počet stran 200, Vydání 1.

Seznam příspěvků

Předmluva

Doktorandské dny jsou tradičním setkáním studentů doktorského studia na Fakultě jaderné a fyzikálně inženýrské ČVUT v Praze. Doktorandi oboru Matematické inženýrství zajišťovaného katedrami matematiky, fyziky a softwarového inženýrství na nich prezentují výsledky své vědecké práce, jejichž tematika pokrývá všechny oblasti aplikované matematiky. Letošní ročník je již čtrnáctým vydáním workshopu, koná se ve dnech 15. a 22. listopadu 2019.

Za materiální podporu děkujeme katedře matematiky FJFI a grantu Studentské vědecké konference SVK 28/19/F4.

Editoři

Runtime Molecular Simulation Nucleation Criterion for Metastable States

David Celný

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

advisors:

Jiří Kolafa, Department of Physical Chemistry, UCT in Prague

Jadran Vrabec, Department of Thermodynamics
Technische Universität Berlin

Roland Span, Department of Thermodynamics, Ruhr-Universität Bochum

Abstrakt. Diese Arbeit beschäftigt sich mit Dampf-Flüssigkeit-Phasenübergängen mittels molekularer Simulationsmethoden. Dabei liegt der Fokus auf Untersuchungen der metastabilen Zustände des Systems, in denen Nukleation stattfindet. Diese metastabile Region befindet sich zwischen den Binodalen und den Spinodalen und ist experimentell nur sehr beschränkt zugänglich. Daher müssen experimentelle Ergebnisse für thermodynamische Stoffdaten in dieser Region durch anderen Methoden ergänzt werden. Diese Problematik wird hier mithilfe von Molekularsimulationen1 und entsprechenden Entscheidungskriterien für die Clusterbildung untersucht. Clusterkriterien können indizieren, ob sich ein molekulares System in einem homogenen Zustand (metastabile Flüssigkeit oder metastabiler Dampf) befindet und die untersuchten Mikrozustände daher zulässige Konfigurationen sind. Integriert werden zwei verschiedene Algorithmen für den Dampf- und Flüssigkeitsteil2 der metastabilen Region auf Basis der unterschiedlichen Dichten beider Phasen. Die aus Simulationen gewonnenen Daten werden analysiert und als weitere Datenquelle für die Entwicklung von Multiparameter-Zustandsgleichungen mit nur einem Maxwell-Loop genutzt. Zahlreiche in der Literatur verfügbare hochgenaue Fundamentalgleichungen verhalten sich aufgrund fehlender Messdaten im metastabilen Bereich physikalisch nicht korrekt. Mit dieser Studie soll dazu beitragen werden, die qualitativen Eigenschaften moderner, vielparametriger Zustandsgleichungen zu verbessern und deren Anwendbarkeit im Nassdampfgebiet zu verbessern.

Stichworte: Metastabile Zustende, Molekulardynamik, Kluster Kriterien

Abstract. This work deals with vapour-liquid phase transitions using molecular simulation methods. The focus is on investigations of the metastable states of the system in which nucleation takes place. This metastable region is located between the binodal and the spinodal and is experimentally only accessible to a very limited extent. Therefore, results for thermodynamic property data in this region have to be supplemented by other means. This problem is investigated here with the help of molecular simulations and corresponding decision criteria for the cluster formation. Cluster criteria can indicate whether a molecular system is in a homogeneous state (metastable liquid or metastable vapour) and therefore decide whether the investigated microstates are permissible configurations. Two different algorithms are integrated for the vapour and liquid part of the metastable region based on the different densities of the two phases. The data obtained from simulations will be analysed and used as a further

data source for the development of multi-parameters equations of state with only one Maxwell loop. Numerous highly accurate fundamental equations available in the literature do not behave physically correct due to missing measurement data in the metastable range. The aim of this study is to improve the qualitative properties of modern, multi-parameter equations of state and to improve their applicability in the wet steam area.

Keywords: Metastable state, Molecular dynamics, Cluster criterium

Abstrakt. Obsahem této práce je zkoumání fázových přechodů mezi parou a kapalinou za použití nástrojů molekulárních simulací. Speciální pozornost je pak věnována výzkumu metastabilních stavů jakožto podmínky nukleace. Samotný metastabilní stav je pak definován jako oblast mezi binodálou a spinodálou jenž je experimentálně velmi obtížná oblast. Z toho důvodu je zapotřebí termofyzikální vlastnosti zajistit pomocí jiných zdrojů. V této studii je použito nástroj molekulárních simulací a patřičného rozhodovacího kritéria, které sleduje formování shluků (Clusterů). Tato cluster kriteria poukazují na to, zda se sytém nachází v homogenním stavu (metastabilní kapalina a nebo metastabilní plyn) a tedy jsou-li generované mikrostavy stále přípustné konfigurace. K tomuto účelu jsou v práci obsažena dvě odlišná kritéria pro kapalnou a plynnou metastabilní oblast, které se od sebe liší svou hustotou. Získaná data budou dále analyzována a náseldně použita jako zdroj pro vývoj multi-parametrických stavových rovnic s pouze jednou Maxwellovou smyčkou. Mnohé v dostupné literatuře vydávají velmi přesné fundamentální stavové rovnice ne zcela přesné předpovědi právě z důvodu nedostatku dat z metastabilní oblasti. Cílem této práce je proto zlepšit kvalitativní vlastnosti moderních multi-parametrických stavových rovnic a zvýšit tak jejich použitelnost v oblasti mokré páry.

Klíčová slova: Metastabilní stav, Molekulární dynamika, Cluster kriterium

1 Introduction

The aim of this research is motivated by the long standing problem of multi-parametric equation of state, namely it is the imprecise prediction of experimentally challenging areas. The problematic is summarized as a multiple Maxwell-loop problem. To resolve such issue the new data sources can be used to access the problematic areas. This study focus specially on the metastable region of phase diagram for selected vapor-fluid liquids. This study utilizes the molecular simulation tools to provide an insight into the behaviour in said region. With the developed method a new dataset can be utilized for nucleation research and also immediately for a new multi-parametric equation of state development.

2 Theoretical background

The theoretical background of the presented work consist of two main areas: nucleation process and molecular simulation. The nucleation extends previous description for metastable state condition investigation. This is an important part of the ongoing research as metastable state is precondition for nucleation. The second portion contain the molecular simulation applied to the metastable systems. This has implication into how the simulation is performed and the modification details are given in this section.

2.1 Nucleation

One of the accepted description of nucleation is as the process of first order phase transition. The phase transition can wary greatly (i.e. phase separation, precipitation. . .) depending on the considered systems. This study therefore consider the vapour $\leftrightarrow$ liquid transition. Motivation for such selection is the continued research of thermophysical fluid properties. An example of considered transition could be the formation of water droplets in clouds for vapour $\rightarrow$ liquid direction. Or carbon dioxide bubbles formation in carbonated drinks in case of vapour $\leftarrow$ fluid direction.

The fundament of following description was presented in context of the Classical Nucleation Theory. The CNT for short was according to Kalikmanov [4] described in first half of twentieth century by Volmer & Weber, Becker & Döring and Zeldovich [13, 2, 14]. Later research extended on the initial ideas but for the purposes of this study the CNT view of metastability is sufficient.

In the framework of considered CNT model this study aims to describe homogeneous nucleation. The term homogeneous imply that no external forces or system impurities(external agents) are present. Consideration of impurities is above the scope of this study.

2.1.1 System stability

System stability is term not only restricted to the field of thermodynamics. A simple mechanical description is for example moving ball on a curved surface. We can observe the ball and denote the its significant states during the observation. The situation can be described by the a potential field (dictated by the curved surface) and the force involved in driving the ball would arise from exchange between potential and kinetic energy of the ball. Observing the stability situations will yield three significant scenarios: When the ball is moving on the curved slope, When the ball is stuck in the small hole for a while but can escape and lastly When the ball has found an absolute minimum of the curved surface. The three states are referred to as: unstable, metastable and stable respectively.

This mechanical conclusion can be also used for thermodynamics. In this scenario the ball is replaced by the thermodynamic system. Curved surface (potential energy) is replaced by the free energy of the system. And the driving force is actually chemical potential. Stability scenarios are then equivalent and system state can be found in the said three states. Similar to the ball the system is also trying to find the most favourable free energy minimum. The word try is significant here because as the ball the system can be stuck in the free energy local minimum. This state is referred to as metastable system and commonly the systems under such conditions are called metastable systems.

All three states are found in 1. In this figure the situation of reaching stable state from metastable state equals to crossing the energy barrier. In the case of ball the accumulated kinetic energy would push the ball over the hill. In the case of thermodynamic system another source of energy apart steep chemical potential gradient is present. The system also has an entropic contribution to the energy. It is therefore possible that with a correct sequence of microstates a barrier can be crossed. In the analogy it could be that the ball is naturally randomly shaking therefore it is possible that it can shake itself over the hill.

The closer observation in thermodynamic sense is given by the phenomenological

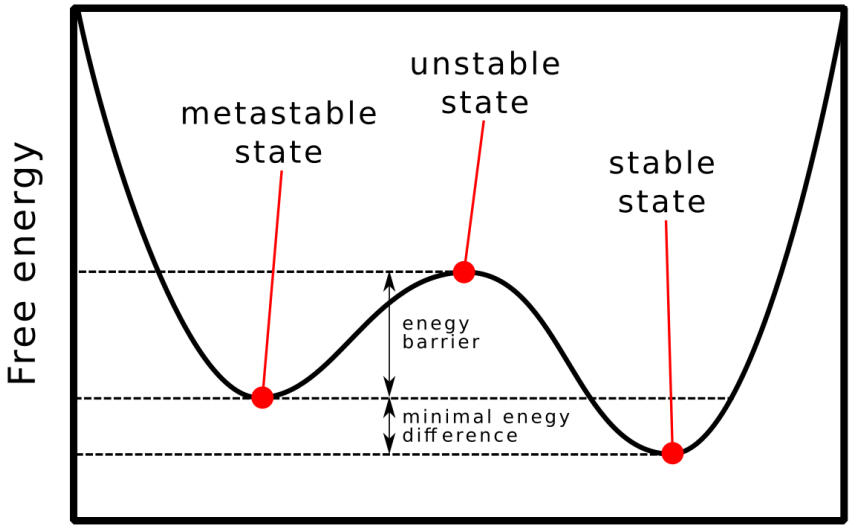


Figure 1: Illustration of free energy function depicting the free energy corresponding to the system stability. This picture was based upon Fig. 3.1 in [4].

thermodynamics. For the thermodynamic system of one fluid with known thermodynamic properties a phase diagram 2 can be constructed. In this picture multiple isotherms were calculated for hypothetical Lennard-Jones fluid into a pressure-density$(\mathrm{p}-\rho)$ diagram. In this diagram two important lines were also drawn, namely binodal and spinodal. These curves are important for the next discussion therefore a simple explanation follows.

Phenomenological thermodynamics summarises binodal as coexistence curve. It is a curve that a phase separation can first occur considering sufficient time. This curve is characterised consequently by condition where the driving force for transition is zero and therefore for phase transition to complete an infinite time would be required.

$$\mu_{\mathrm{vapor}}(\mathrm{p},\rho) = \mu_{\mathrm{liquid}}(\mathrm{p},\rho) \tag{1}$$

This leads to the next line. For spinodal the chemical potential gradient increasingly favour the phase transition. This character proceed until even an infinitesimal fluctuation would lead to the phase transition. The criteria can be again summarised that the energy barrier will lower itself into a saddle point. Or equivalently that the second derivative of the free energy equals to zero (surrounding free energy slopes sign is enough to conclude the saddle point existence).

These two curves play a pivotal role in stability for thermodynamic system because binodal separate stable from metastable states. And spinodal separate metastable from unstable states. In following section more attention is given to the metastable state.

2.1.2 Metastable state

The previous paragraphs showed that the metastable state is not of permanent nature. Given infinite time the state would collapse eventually into the global minimum.

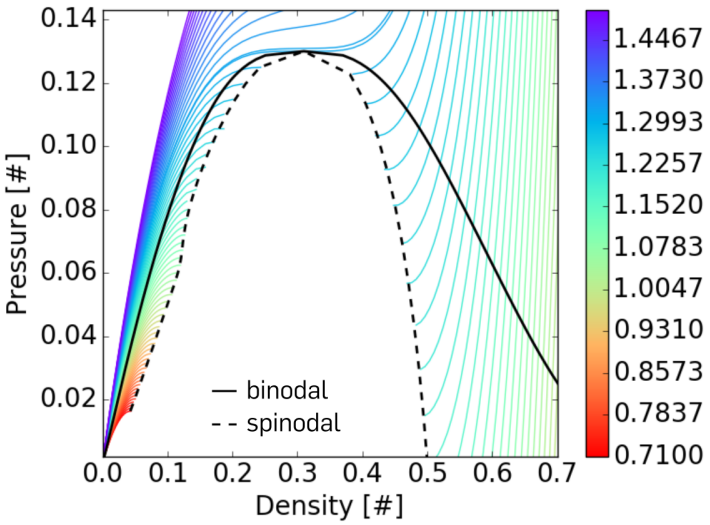


Figure 2: Illustration of system stability at the p, ρ phase diagram.

Therefore it is interesting to discuss what influence the time required for the state to cross the energy barrier. It was already mentioned that the nature of transition is stochastic in principle. It is to be expected therefore that for larger system the event of creating suitable microstates will be more probable. In principle these suitable microstates contain precursors of the local phase change. Depending on the transition direction we refer to these precursors as clusters for vapour $\rightarrow$ liquid or voids in case of vapour $\leftarrow$ liquid transitions. These precursors are further collectively referred to as local density inhomogeneity. This notation help to simplify description and is a significant realization in the latter simulation section.

$$J = J_0 \exp\left(\frac{-\Delta G^*}{k_\mathrm{B} T}\right) \tag{2}$$

In theory all factors that contribute to the nucleation rate J described in previous study have their consequence in dynamics of local density inhomogeneity. The system size is for example relevant because of the J_0 pre-factor. Next influencing factor is the Gibbs energy gradient. This correlates to the height of energy barrier and can be interpreted as how far the system is from binodal (or alternatively how close the system is to spinodal). The energy barrier is also influenced by the system temperature as it can be inferred from 2 observing the decrease in width of metastable region for increase in temperature. Theoretically it is clear because higher temperatures translates into system with more accumulated energy (similar to the shaking ball analogy with more vigorous shaking).

From the previous it can be concluded that the system exhibit a dynamic behaviour. The system is furthermore more likely to exhibit the density inhomogeneity the closer it is to spinodal. And the lastly the overall situation drastically differ between the metastable liquid and metastable vapour regions. The problem basically changes from the search for groups to the search for empty spaces. In the developed method these challenges has to be addressed for meaningful results to be obtained.

2.2 Molecular simulations

An alternative approach to experiment are molecular simulation tools performing a pseudo-experiment in computer memory [6]. This offers a benefit of a reproducible system evolution which can be performed in experimentally unfeasible setting or measure experimentally unreachable properties. This advantage is also a partial drawback as the experiment reflects the nature while a simulation with incorrect setup provide unphysical results. In this regard the simulation has to be carefully regarded as well as interpretation of the produced results. It is additionally very important to validate the simulations with existing models and cross-correlate.

The majority of properties obtained from simulation tool are in the form of a mean values. When different thermodynamic ensembles are considered one can obtain the mean energy of the system, pressure and local density distribution. In consideration of the dynamic character of the problematic this study use the Molecular dynamics(MD). It is noted that Monte Carlo(MC) can be used as well and it is task for future to cross-validate both methods.

2.2.1 Molecular dynamics

Molecular dynamics considers a Newtonian description of a mechanical system with the prescribed potential field. In this setup, the modelled system gets its mechanical behaviour from the Newtonian physics although methods taking into account quantum behaviour are also available. More detailed explanation can be found in [3, 6, 1].

The general idea of molecular simulation is that for a provided initial configuration the model produce time series of snapshots representing a state of system in simulated time. These time snapshots can be used for post-processing, or quantities of interest can be computed during the simulation. For example the immediate energy of the system can be computed online. Analysis of these output data can yield a desired mean property value. The analysis of time snapshots produces also structural information about the system such as the radial distribution function or cluster distribution.

Expanding further on the nature of the Newtonian problem are the interactions. The simple form of potential of non-bonded (denoted as U_{nb}) system (system containing only one type of unbounded atoms such as Ar) is

$$U_{nb}(r^N) = \sum_{i=1}^{N} u^E(r_i) + \sum_{i=1}^{N}\sum_{j>i}^{N} u^P(r_i, r_j) + \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{k>j}^{N} u^T(r_i, r_j, r_k) + \dots \qquad (3)$$

where the first sum represents external effecting potential u^N, the second sum stands for two-body interaction and following sums similarly represents the k-body interactions. In many cases the external potential is zero. Additionally, considering the ratio of two-body, three-body and further interactions, the three of more interaction terms are to be neglected without causing significant decrease of model accuracy. The two-body potential can be reduced from two position vector to a function of interparticle separation, $u(r_{ij})$.

The corresponding force per particle is obtained as negative differentiation of the potential

$$U(r^N) = \sum_{i,j} u(r_{ij}) \tag{4a}$$

$$f_i = -\frac{\partial}{\partial r_i} U(r^N) = -\sum_{i \neq j} u'(r_{ij}) \frac{r_{ij}}{|r_{ij}|}. \tag{4b}$$

The force formula is evaluated directly instead of run-time differentiation. The force per particle is then used to calculate atom velocities and new atom positions for a time advanced by a constant time-step. In the case of bare calculation the computational time is then incremented leading to the next iteration.

In the basic simulation settings following key performance points are noted. The greatest time requirement is imposed by the potential/force calculation as could be inferred from the form of equation (4b), which is a $\mathcal{O}(n^2)$ problem. The remainder of calculation is in form of $c\,\mathcal{O}(n)$ where the constant c represent the additional overhead in position updates and inclusion of the bond constrain solver of no greater complexity than $\mathcal{O}(n)$. Avoiding the issues of file generation and incurred slowdown the simulation can still be significantly delayed by the local density detection criterion.

2.2.2 Detection criteria

The first thought about the detection of the density inhomogeneity was targeted on the detection of clusters. Reason for this is the clearer geometrical interpretation of the task. The task can be formulated with the use of graph theory as: detection of connectivity components that exceed a given vertices count. Generally this also means calculation demand leading to methods of $\mathcal{O}(n^2)$ complexity. The criterion main purpose is fulfilled when the system reach the predefined threshold and the simulation is terminated after the mean values are returned. The thresholds (both cluster and molecules per cluster) are determined by the user and therefore the desire is for the criteria to be robust in this input.

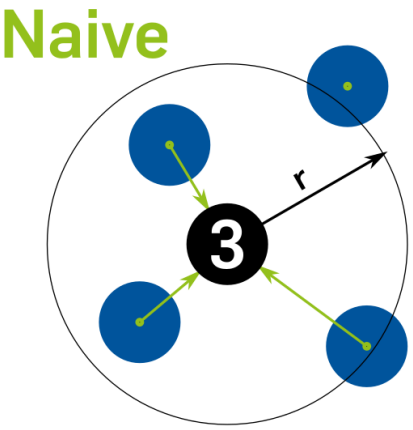


Figure 3: Naive cluster criteria with the outline of its functionality.

In the illustration 3 explanation of what simple distance check means for single molecule is given. The straightforward nature of the criteria mimics for example the

Verlet list method. Naive criteria provide an easily parallelized but imprecise method of detesting groups of molecules. Unfortunately the method is very sensitive in the choice of the radius r 4σ. This result in varying success of usage and significant prior knowledge demand how to setup the criteria parameters. With larger radii another problem becomes apparent with over-counting of the cluster. It is easy to imagine that one physical cluster is being reported from multiple molecules in its centre (as they also fall into prescribed count). This is a significant issue that prevents a simple usage of the criteria as the demand for system prior knowledge is too great.

With the previous method drawback in mind a precise cluster counting criteria was adopted. Based on the formulation of Jr. Stillinger [11] the criterion with precise but more calculation intensive approach was designed. The principle is again illustrated at 4, where the recursive nature of the criteria is depicted by the cascade of the green arrows as they add up to the originating molecule neighbour count. The recursive search has its drawback in the fact that it is hard to parallelize and therefore is unfeasible for some applications. The benefit of the method is that it provide a precise number of clusters that can be of arbitrary shape.

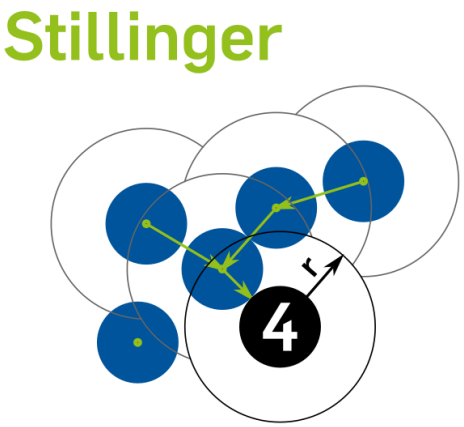


Figure 4: Stillinger cluster criteria with the outline of its recursive nature.

Stillinger criteria is a robust method but unfortunately valid only for clusters. As the part of the research task also liquid case criteria is required. The task is to obtain the benefit of both naive criteria in the simplicity and parallel design in combination with the benefit of precise count of clusters and voids later on. This complicated task was resolved by the grid criteria. The original idea of grid was inspired by the M. Horsh kd-tree space partitioning algorithm. In this study a simpler method of regular grid was chosen because of the efficiency concerns. The figure 5 illustrate the method with the imaginary grid in black colour. The neighbour molecule counts are made inversely from grid and the values in the grid-points are tested in the decision. This can speed the whole calculation as the iteration can utilize the regularity of the grid as well as truncate the far grid planes. The grid also is easy to parallelize and mitigate the over-counting to maximal factor of 8 (the amount of grid points sharing the same voxel). The problem is already further mitigated by significantly smaller radius that is comparable to Stillinger method and vary around $r = 1.5\sigma$.

This grid method is the culmination of the development process where multitude of variation of previous were considered. Developed criterion fulfils the initial criteria and also satisfy the demand for unified criteria that can easily switch between the two transition states. This feature alone is beneficial as it allows for robust criteria with potential for further optimization and extension for example into automatic phase transition type independent criterion.

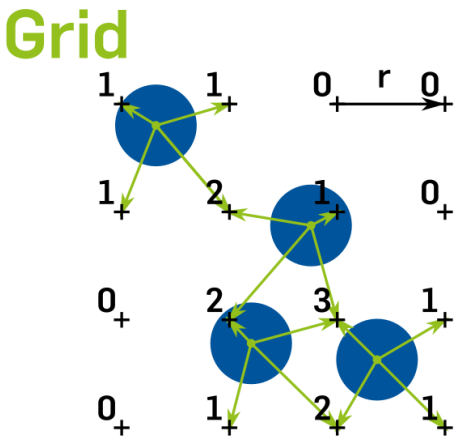


Figure 5: Grid cluster criteria with the outline of its inverse counting.

3 Solution method

In following section a method is illustrated that is used for the preliminary result calculation. The method relies heavily on the external tools cooperation, therefore the necessary intercommunication is discussed here. This section gives further detail on the external tools used as well as the construction of metastable state condition show in the 6.

TREND The first package is called TREND currently in version 4 [9]. The following description is the package summary from the program manual:
This software package has been developed as part of the research in thermodynamic property models of the thermodynamics group at the Ruhr-University Bochum. It is written in FORTRAN 95, using the Intel Visual Fortran compiler XE (Version 12 or higher) for debugging and compilation. It is subject to ongoing development and will incorporate extended functionality in future versions. The thermodynamic properties are mainly calculated using highly accurate equations of state (EOS) explicit in the Helmholtz free energy. More information on the general structure of Helmholtz-type equations of state can be found in [10]. For mixtures, the non-ideal interaction of fluids is accounted for using mixture models in the way they were developed for the GERG-2004 EOS [5]. Other equation models, e.g., PR, SRK, or LKP, are also available.

ms2 The second package is called ms2 [7] presently in version 3. The following description is the package summary from the web page of the program:

The molecular simulation program ms2 is designed for the calculation of thermodynamic properties of bulk fluids in equilibrium. ms2 features the two main molecular simulation techniques, molecular dynamics (MD) and Monte-Carlo. It supports the calculation of vapor-liquid equilibria of pure fluids and multi-component mixtures described by rigid molecular models on the basis of the grand equilibrium method. Furthermore, it is capable of sampling various classical ensembles and yields numerous thermodynamic properties. To evaluate the chemical potential, Widom's test molecule method and gradual insertion are implemented. Transport properties are determined by equilibrium MD simulations following the Green-Kubo formalism. ms2 is written in Fortran90 and optimized for a fast execution on a broad range of computer architectures, spanning from single processor PCs over PC-clusters and vector computers to high-end parallel machines. The standard Message Passing Interface (MPI) is used for parallelization and ms2 is therefore easily portable onto a broad range of computing platforms. Feature tools facilitate the interaction with the code and the interpretation of input and output files. The accuracy and reliability of ms2 has been shown for a large variety of fluids in preceding work.

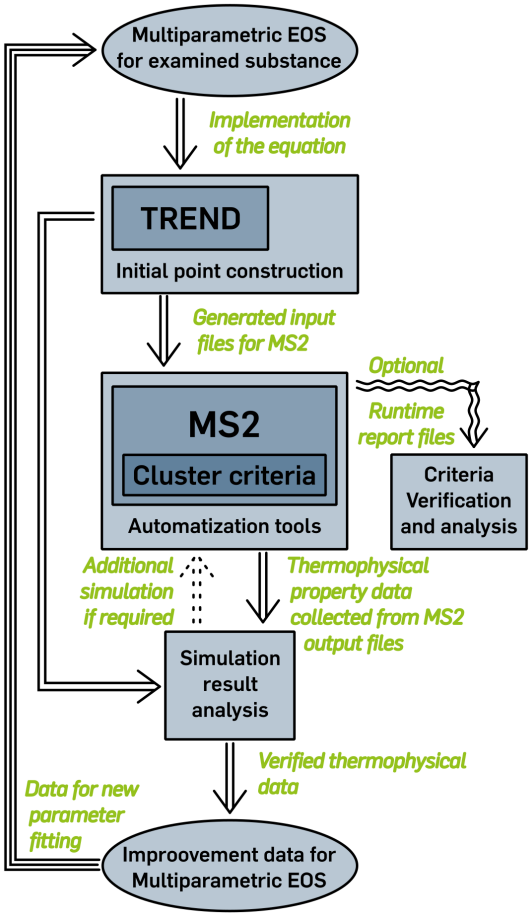


Figure 6: Overview of the method design with the incorporated external tools TREND & ms2.

The initial metastability condition of the investigated system were obtained by performing a EOS calculation with the *TREND* package. The required surrounding

tools were developed for the detection of binodal spinodal region and visualization. Once the metastable region is identified based on the binodal, spinodal construction the system condition can be sampled. This step actually consist of selecting the densities on desired isotherms in $p - \rho$ phase diagram. The points themselves are further exported to format readable by the *Automatization tools* surrounding the *ms2* package. This package is responsible for performing the MD simulation of the system. The system settings and input file generation and management are part of developed *Automatization tools*. This design ensures the process can proceed smoothly even with the remote calculation on supercomputer.

The core part of *ms2* that was also developed by the author is the detection criteria (*Cluster criteria* in the figure 6). The criteria are responsible for detecting the local density inhomogeneity and stopping the simulation. That ensures that the data calculated are still of homogeneous system. Functionality of criteria were cross-validated by external evaluation of the generated report files as well as consistency checks with available EOS prediction. The prime check was performed for artificial LJ fluid for which the precise EOS is known to work even in metastable region.

With the proper set up, the calculated results are send to the analysis program. In the program the basic data are visualised and compared with the EOS prediction. With the analysis and dataset consistency check the final values can be further utilised. One of the immediate application lies in the new multi-parametric EOS creation. Further application include the nucleation research or even investigation of metastable state in general.

4 Intermediate results

This study summarises results of the research done as a part of exchange program Erasmus. During the stay in research group of R. Span the application of the research was the new data set for multi-parameteric EOS. In this regard the results currently available are tailored to that requirement even though the research is not restricted only to that application. The investigated fluids were hypothetical LJ fluid, Oxygen and Nitrogen. These fluids are relevant to the Helmholtz energy EOS research and the current focus of the eam in Bochum.

In accordance to the described method the system evolution is presented first to illustrate the functionality of the grid criteria itself. It is to be noted that the set of figures 4b depict the grid values of the simulated LJ fluid system with 864 particles and $12x12x12$ grid. The system condition correspond to the metastable vapour and therefore a droplet formation can be observed. The parameters for simulation were set to $1.5\sigma = 1.5$ for grid width and cluster is considered when six of more molecules are found in the neighbourhood of the grid point. This alone would result in too harsh of a criterion as the random fluctuation can result in a such a group even in stable system. Therefore the simulation was stopped only after 1.5% of the grid reported a density inhomogeneity. Multiple frames with the detected clusters are shown in the figure from the beginning portion of simulation. The number of clusters is still low here an the cluster kinetics can be easily observed, as shown in frames 10-12. here the cluster dissolve (f10-f11) and new formation/movement of the cluster (f11-f12) occurs in between of 10 iteration steps and

0.025 ps.

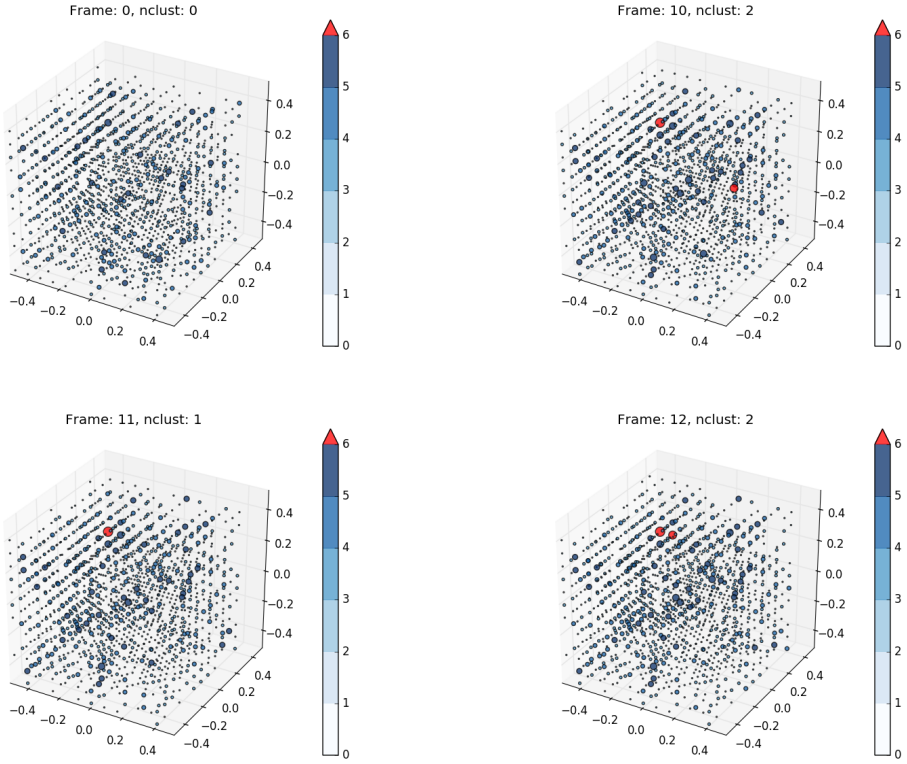


Figure 7: Grid cluster criteria output for LJ fluid at reduced temperature of 1.219677 and reduced density of 0.142332.

The validity of the method was verified on the example of LJ fluid where the EOS is known to provide good description even in metastable system. Therefore for example pressure not used as input to the MD simulation should be sufficiently close to the one sampled from EOS. It can be seen in 8 that very close reproduction was achieved not only for the stable region of LJ fluid but more importantly for the metastable region. The comparison shows the performed sampling on both sides of phase diagram with symbols separated to circles for vapour and squares for liquid and stars corresponding to the simulated data.

The sampling itself contain 147 points divided into vapour side in the left part of figure and liquid part on the right. The left figure was executed with vapour grid criteria type searching for more than 8 molecules in neighbourhood (the rest of criteria parameters were same) and the liquid side searched for voids that have 1 or less molecules in neighbourhood. These parameter setting were found to provide best representation.

These results are significant. Not only because it proves that the method works as intended but it can be used further on more complicated systems such as real fluids. The second benefit is that as the criteria employs scaling to the sigma distance, therefore the discovered parameters can be reused for the simple real fluids as well (not large multi-atomic molecules).

The same method was further applied to the pure oxygen and nitrogen. For oxygen

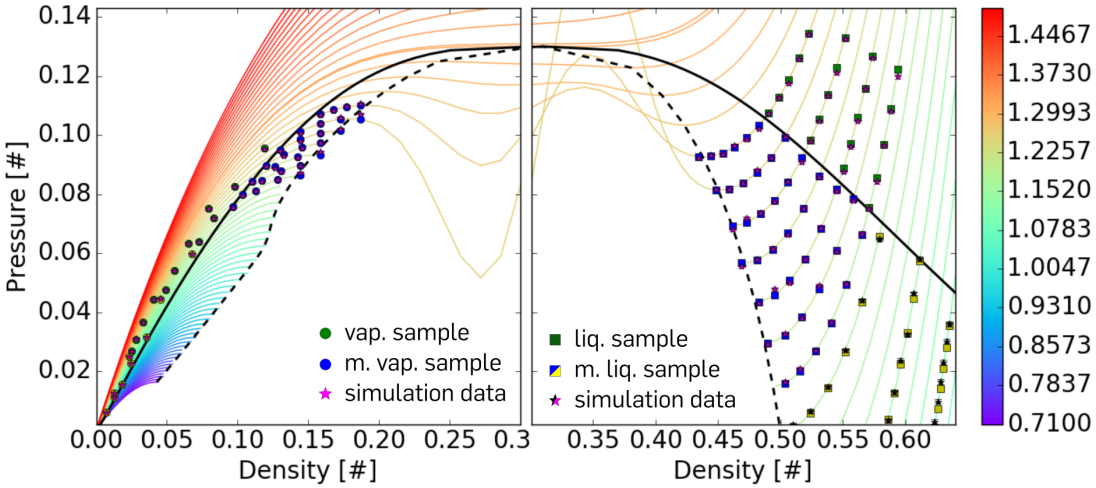


Figure 8: Comparison of the LJ fluid sampling, simulation data and EOS prediction by Thol *et.al.* 2016 [12]

the reference Helmholtz multi-parametric EOS by Schmidt & Wagner (1985) [8] was used. Similarly as the Thol *et.al* in figure 8 and other multi-parametric EOS it has a problem of multiple Maxwell loops. This issue is characterised by the oscillation of the isotherms in the metastable and primarily the unstable region of the phase diagram. The equation therefore provide imprecise predictions in said regions as no data are available to correct that and stable fluid data are not applicable. The research done in last year should help to remedy the situation but there are also speculation if the cubic EOS can be used instead (as cubic EOS does not suffer from multiple Maxwell loops). Because of it the Peng-Robinson(PR) cubic EOS was used for comparison in case of Nitrogen to observe how the discrepancies behave.

When the figure 9 is presented the immediate notion is that the prediction is incorrect. That is also the case because the EOS is initially valid only at the stable region. At that region the simulation align well with the EOS. But when the metastable region is considered the discrepancies are clearly visible. This is an opportunity for the simulation data to help because as it was mentioned earlier the Maxwell-loop issue is present and that influence the EOS validity in metastable region. What is the interesting realization is that the simulation data have the correct shape of isotherm and also that the the trend is consistent with the thermodynamic knowledge (i.e. the steepness of the slope from liquid side and minimum values temperature dependence). The data also illustrate the shift in spinodal curve from violet line to the a new position that connect the minimum of simulated isotherms. The left side of the graphs behaves in accordance to the expectation because the predictions metastable vapour are less problematic.

Observing the next figure 10 in similar detail it is interesting to notice that the simple PR EOS has already imprecise prediction in the stable liquid region. This quantitative shift continue also into metastable liquid area. The prediction of the spinodal and the character of the isotherm there is better but quantitative difference is still present. With this data is is clear to see that using cubic EOS can alleviate some issues in metastable regions but price has to be paid in stable liquid as was already widely accepted in

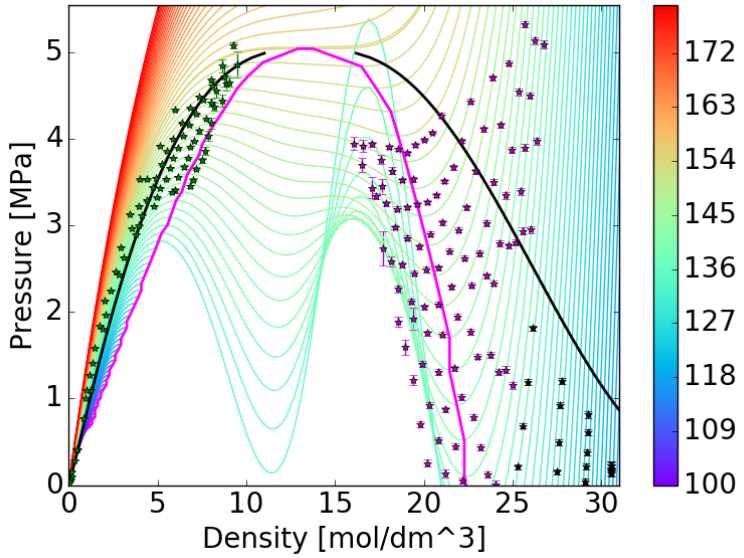


Figure 9: Comparison of the simulation data for oxygen and EOS prediction by Schmidt & Wagner 1985 [8]

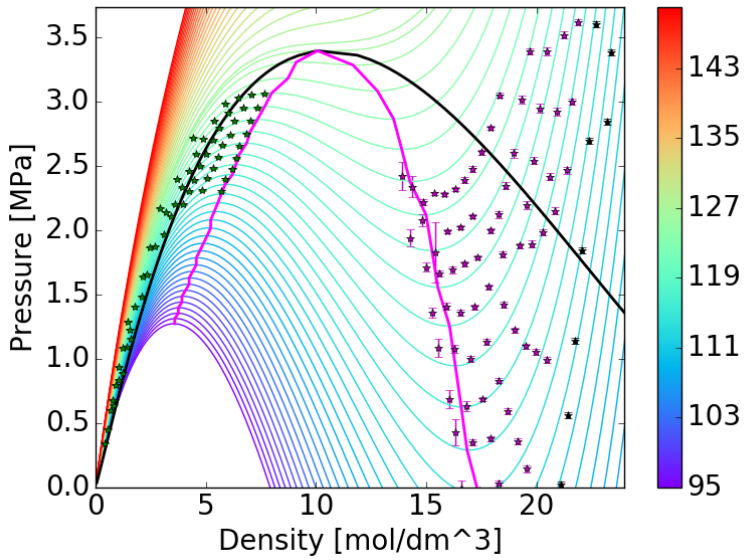


Figure 10: Comparison of the simulation data for nitrogen and PR EOS prediction

Equation of state community. In this study the data finally provide further insight into the problematic with potential to help more in the future. The vapour side exhibit only minor discrepancies.

5 Conclusion

Presented study is an overview of the work performed in the span of last year on the ground of Ruhr Universität Bochum. The task of the research was to investigate the metastable region of phase diagram with molecular simulation tools and provide a reliable data for a new EOS that would provide better prediction in this region. For this reason the metastability was theoretically investigated and a relevant method was later developed for the metastable system properties calculation. This method utilize the molecular simulation tools provided in external software package. The software was examined in great detail and specialised detection criteria was implemented such that the simulation results can be used also for metastable system simulation. The overarching supporting tools were implemented to facilitate external packages functionality and validation. The method was verified on the LJ fluid yielding in very precise agreement. Method was further used on two pure fluids: Oxygen and Nitrogen. The simulated data provided new unique insight into the metastable region. Primarily the metastable liquid region was investigated. With these result the foundation of the future cooperation and further research have been laid with rewarding application in both primary research and industrial applications.

References

[1] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 2 edition, (2017).

[2] R. Becker and W. Döring. *Kinetic treatment of the formation of nuclei in oversaturated steam*. Ann Phys **5** (1935), 719–752.

[3] D. Frenkel and B. Smit. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications (Computational Science Series 1)*. 2 edition, (2001).

[4] V. Kalikmanov. *Nucleation theory*, volume 860. Springer, (2012).

[5] O. Kunz, R. Klimeck, W. Wagner, and M. Jaeschke. *The gerg-2004 wide-range equation of state for natural gases and other mixtures*. In 'GERG TM15', volume 6, Fortschritt-Berichte VDI (2007).

[6] D. C. Rapaport. *The art of molecular dynamics simulation*. Cambridge university press, (2004).

[7] G. Rutkai, A. Köster, G. Guevara-Carrion, T. Janzen, M. Schappals, C. W. Glass, M. Bernreuther, A. Wafai, S. Stephan, M. Kohns, S. Reiser, S. Deublein, M. Horsch, H. Hasse, and J. Vrabec. *ms2: A molecular simulation tool for thermodynamic properties, release 3.0*. Computer Physics Communications **221** (2017), 343 – 351.

[8] R. Schmidt and W. Wagner. *A new form of the equation of state for pure substances and its application to oxygen*. Fluid Phase Equilibria **19** (1985), 175–200.

[9] R. Span, R. Beckmüller, T. Eckermann, S. Herrig, S. Hielscher, A. Jäger, E. Mickoleit, T. Neumann, P. S. M., B. Semrau, and M. Thol. *TREND*. Lehrstuhl für Thermodynamik, Universitätsstr. 150, 44801 Bochum, Germany, (March 2019). Version 4.0.

[10] R. Span. *Multiparameter equations of state: an accurate source of thermodynamic property data*. Springer Science & Business Media, (2013).

[11] F. H. Stillinger Jr. *Rigorous basis of the frenkel-band theory of association equilibrium*. The Journal of Chemical Physics **38** (1963), 1486–1494.

[12] M. Thol, G. Rutkai, A. Köster, R. Lustig, R. Span, and J. Vrabec. *Equation of state for the lennard-jones fluid*. Journal of Physical and Chemical Reference Data **45** (2016), 023101.

[13] M. Volmer and A. Weber. *Nuclei formation in supersaturated states*. Z. Physik. Chem., ll9 (1925), 277–301.

[14] Y. B. Zeldovich. *On the theory of new phase formation: cavitation*. Acta Physicochem., USSR **18** (1943), 1.

Short-Time Fractal Analysis of Biological Autoluminescence*

Martin Dlask

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

advisors:

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

Pavel Sovka, Department of Circuit Theory
Faculty of Electrical Engineering, CTU in Prague

Abstract. Biological systems manifest continuous weak autoluminescence, which is present even in the absence of external stimuli. Since this autoluminescence arises from internal metabolic and physiological processes, several works suggested that it could carry information in the time series of the detected photon counts. However, there is little experimental work which would show any difference of this signal from random Poisson noise and some works were prone to artifacts due to lacking or improper reference signals.Here we apply rigorous statistical methods and advanced reference signals to test the hypothesis whether time series of autoluminescence from germinating mung beans display any intrinsic correlations.Utilizing the fractional Brownian bridge that employs short samples of time series in the method kernel, we suggest that the detected autoluminescence signal from mung beans is not totally random, but it seems to involve a process with a negative memory. Our results contribute to the development of the rigorous methodology of signal analysis of photonic biosignals.

Keywords: short time series, random process, fractional Brownian bridge, chemiluminescence, biological autoluminescence

Abstrakt. Biologické systémy vykazují nepřetržitě slabou autoluminiscenci, která je přítomna i bez vnějších podnětů. Jelikož autoluminiscence vzniká z vnitřních metabolických a fyziologických procesů, několik prací dříve naznačovalo, že by mohla nést informaci v časové řadě detekovaných počtů fotonů. Existuje však jen málo experimentálních prací, které by ukázaly jakýkoli rozdíl tohoto signálu od náhodného Poissonova šumu a použité metody byly náchylné vůči přítomnosti artefaktů kvůli chybějícím nebo nesprávným referenčním signálům. V této práci používáme statistické metody a pokročilé referenční signály, abychom zjistili, zdali řada autoluminiscence z klíčících fazolí vykazuje nějakou vnitřní korelaci. S využitím zlomkového Brownova mostu, který využívá krátké vzorky časové řady, bylo zjišteno, že detekovaný autoluminiscenční signál z fazolí mungo není náhodný, ale, jedná se o proces s negativní pamětí. Naše výsledky s použitím rigorózní metodologie jsou novinkou v oblasti analýzy fotonických biosignálů.

Klíčová slova: krátké časové řady, náhodné procesy, zlomkový Brownův most, chemiluminiscence, biologická autoluminiscence

*This work has been supported by the grant SGS17/196/OHK4/3T/14

Full paper: M. Dlask, J. Kukal, M. Poplová, P. Sovka, M. Cifra M. *Short-time fractal analysis of biological autoluminescence.* PLOS ONE 14 (2019), e0214427. `https://doi.org/10.1371/journal.pone.0214427`

Boundary Layer Flow Simulations Using Lattice Boltzmann Method*

Pavel Eichler

2nd year of PGS, email: eichlpa1@fjfi.cvut.cz
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. In this contribution, we deal with the numerical lattice Boltzmann method for simulations of a turbulent fluid flow in the boundary layer above smooth and rough walls. First, a short introduction to the boundary layer theory and the lattice Boltzmann method is presented. Next, two benchmark tests are performed. It was observed that LBM produce satisfactory results in comparison with results in the literature. In the second benchmark test, LBM is in a good agreement with a code based on the finite difference method.

Keywords: lattice Boltzmann method, boundary layer, turbulent flow

Abstrakt. V tomto příspěvku se zabýváme numerickou metodou lattice Boltzmann pro simulace turbulentního proudění tekutin v oblasti mezní vrtvy nad hladkou a hrubou stěnou. V první části je krátce uvedena teorie mezní vrstvy a následně metoda lattice Boltzmann. V další části jsou provedeny dva testovací příklady. V případě proudění nad hladkou stěnou, výsledky získané pomocí LBM jsou srovnatelné s výsledky v dostupné literatuře. V případě druhého testu jsou výsledky získané pomocí LBM v dobré shodě s výsledky vypočtenými pomocí kódu založeném na metodě konečných diferencí.

Klíčová slova: metoda lattice Boltzmann, mezní vrstva, turbulentní proudění

1 Introduction

Computational fluid dynamics (CFD) is one of the intensively studied fields in numerical mathematics. In practice, problems involving fluid dynamics can be found frequently, e.g., in determining aerodynamics in the automotive industry, in the aerospace industry, in the energy industry, in health care, and in many other industries. The advantages of numerical simulations include lower financial costs compared to real experiments, time savings, and adaptability to partial problems. CFD consists of several numerical methods for simulating fluid dynamics. These methods can be classified from several points of views. One criterion is the description of the fluid itself. The fluid can be described macroscopically (as a continuous structure), microscopically (as the movement of the individual particles, such as atoms or molecules) or mesoscopically. The mesoscopic

*This work has been supported by the Czech Science Foundation project No. 18-09539S, the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/194/OHK4/3T/14, and by the project NV19-08-00071 of the Ministry of Health of the Czech Republic.

description is based on the kinetic theory and uses a probability distribution function to describe the fluid.

One of the most challenging problems often investigated by CFD is the turbulent fluid flow. Turbulent fluid motion is characterized by chaotic changes in pressure and velocity fields. This type of flow is related to all application fields mentioned in the previous paragraph and, thus, the solution of turbulent fluid flow is extensively investigated.

The turbulent motion consists of several structures at different size scales. Thus, to accurately simulate turbulent flow, a high-resolution mesh is crucial. However, there are some turbulent models which moderate these restrictions on the computational mesh, for example, wall function methods, LES (Large Eddy Simulations) methods, etc.

One of the numerical methods suitable for simulating turbulent flows is the lattice-Boltzmann method (LBM) developed in the 1980s from cellular automata. Later, it was shown, [13], that LBM can be derived directly by discretizing Boltzmann's transport equation, which describes the fluid in mesoscopic sense. It implied the independence of this method from previous cellular automata. The relation to partial differential equations describing fluid dynamics in the macroscopic sense can be shown, for instance, by the asymptotic analysis, see [13].

LBM is suitable for parallel implementation because in the algorithm, what is non-linear is local and what is non-local is linear. Therefore, as hardware for parallel computations develops, LBM develops. Thanks to this development of LBM, there are several submethods that differ mainly in the approximation of the collision operator. In this thesis, we focus on the cumulant LBM (CuLBM) [9, 8].

The inspiration for the experimental benchmarks comes from the collaboration with the Czech Academy of Sciences based on the project of the Grant Agency of the Czech Republic no. 19-09539S entitled Large structures in the boundary layers over complex surfaces in high Reynolds numbers. The results are compared both with results from the literature and from the numerical results provided by Dr. Vladimír Fuka, who developed a finite-difference simulator CLMM.

2 Mathematical model

2.1 Non-dimensional numbers

In the mathematical description, the non-dimensional numbers are used for the description of the problems because the non-dimensionality guarantees the independence of the particular description, i.e., choice of the unit system, coordinate systems, etc. In the fluid dynamic, the Reynolds number Re is often used to express the similarity between different problems. This number is defined as

$$Re = \frac{U_\infty \, d_0}{\nu}, \tag{1}$$

where ν $[m^2\, s^{-1}]$ is the kinematic viscosity, U_∞ $[m\, s^{-1}]$ is the free stream velocity, and d_0 $[m]$ is the characteristic length, [14].

2.2 Equations for incompressible, isothermal, newtonian fluid

The following incompressible Navier-Stokes equations describe the dynamics of Newtonian, incompressible fluid in isothermal domain Ω as, [13],:

$$\nabla \cdot \vec{u} = 0, \tag{2a}$$

$$\rho \frac{D\vec{u}}{Dt} = -\nabla p + \mu \nabla^2 \vec{u} + \rho \vec{g}, \tag{2b}$$

$$p = c_s^2 \rho. \tag{2c}$$

2.3 Boundary layer

Under inviscid fluid approximation, i.e., the assumption of zero dynamic viscosity, the incompressible Navier-Stokes equations (2) reduce to Euler equations

$$\nabla \cdot \vec{u} = 0, \tag{3a}$$

$$\rho \frac{D\vec{u}}{Dt} = -\nabla p + \rho \vec{g}, \tag{3b}$$

$$p = c_s^2 \rho. \tag{3c}$$

It is easier to find the solution of Eqs. (3) due to the absence of the viscous stress tensor and, in some cases, these equations produce satisfactory results in comparison with experiments. The satisfactory results are obtained mainly for high Reynolds numbers, because for fixed mean velocity and characteristic length, as $\mu \to 0^+$, $Re \to +\infty$.

On the other hand, the absence of viscosity leads to no shear stress at the walls and, thus, for instance, Eqs. (3) cause that the no-slip condition is not satisfied at the wall.

For the large Reynolds number flows, the velocity transition from the solution of the Euler equations with a finite non-zero value close to the wall to zero directly at the wall is important in many applications. This transition layer was described in 1904 by L. Prandtl and is called the boundary layer.

With the concept of the boundary layer, the flow region can be divided into two unequally large regions called inviscid outer flow and boundary layer. The inviscid outer flow is located in the bulk flow, where the viscosity can be neglected, and the fluid can be described by Euler equations (3). On the other hand, in the region near the wall, the viscous forces must be taken into account.

One of the most significant parameters at the boundary layer theory is the thickness of the boundary layer δ, defined as the distance from the wall to the inviscid outer flow. According to [14], as Re grows, δ decreases. Thus, to accurately simulate the turbulent fluid flow in the boundary layer, a high-resolution mesh is essential.

Next, in the turbulent layer theory, it is common to operate with non-dimensional quantities. We introduce the dimensionless wall coordinate y^+ as

$$y^+ = \frac{u_\tau y}{\nu}, \tag{4}$$

the shear velocity u_τ as

$$u_\tau = \sqrt{\frac{\overline{\tau_w}}{\rho}}, \tag{5}$$

where $\overline{\tau_w}$ is the time-averaged wall shear stress, and the non-dimensional velocity u^+ as

$$u^+ = \frac{u}{u_\tau}. \tag{6}$$

It is common to use rather the Reynolds number Re_τ instead of Re, where

$$Re_\tau = \frac{u_\tau L}{\nu}. \tag{7}$$

3 Numerical method

Let the computational domain Ω be rectangular and discretized by a regular lattice $\hat{\Omega}$ [10].

3.1 Introduction to lattice Boltzmann method

The Lattice Boltzmann method (LBM) is a numerical method which is based on the mesoscopic description of the fluid using discrete distribution functions, see [13]. Although LBM deals with the mesoscopic description, it can be shown that LBM can be used to solve partial differential equations such as, for instance, the Navier-Stokes equations, [15]. One of the advantages of the LBM is that it lacks the need for solving the Poisson's equation for the pressure-velocity correction (the pressure is described by the equation of state), which is usually the most time-consuming procedure of the standard CFD solvers.

The fluid evolution in the computational domain is described by discrete probability density functions $f_{ijk}(x, y, z, t)$, where the subscripts $i, j, k \in \{-1, 0, 1\}$ correspond to the non-dimensional discrete velocities, $x, y, z \in \mathbb{R}$ are the spatial coordinates, and t is the time. According to the discrete velocity model, we have different sets of discrete microscopic velocities $\vec{\xi}_{ijk} = (ic, jc, kc)^T$, where $c = \frac{\Delta x}{\Delta t}$ is the non-dimensional lattice speed, Δx is the non-dimensional distance between neighboring lattice sites, and Δt is the non-dimensional time step. In this work, we focus on the velocity model D3Q27 only, i.e., in 3D, we use 27 discrete microscopic velocities $\vec{\xi}_{ijk}$, see Fig. 1.

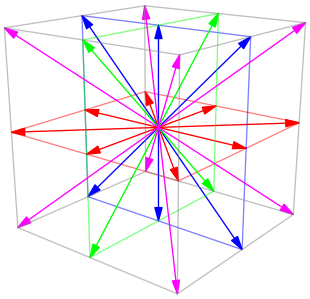


Figure 1: D3Q27 velocity model for $\vec{\xi}_{ijk}$, $i, j, k \in \{-1, 0, 1\}$.

The evolution of the system is given by the discrete Boltzmann equation

$$f_{ijk}(x, y, z, t + \Delta t) = f_{ijk}(x, y, z, t) + \mathcal{C}_{ijk\,xyz}(t), \tag{8}$$

where $x, y, z \in \hat{\Omega}$, $t > 0$, and $\mathcal{C}_{ijk\,xyz}(t)$ is the discrete collision operator.

The macroscopic quantities ρ and $\rho\vec{u}$ describing the fluid can be recovered as:

$$\rho = \sum_{i,j,k\in\{-1,0,1\}} f_{ijk} \tag{9a}$$

$$\rho\vec{u} = \sum_{i,j,k\in\{-1,0,1\}} f_{ijk}\,\vec{\xi}_{ijk} + \frac{\Delta t}{2}\vec{g} \tag{9b}$$

LBM has rapidly evolved during the last decades and several submethods emerged, such as SRT-LBM (LBM with single relaxation time) [10], MRT-LBM (LBM with multiple relaxation time) [2], CLBM (Cascaded LBM) [6], FCLBM (factorized LBM) [7], CuLBM (Cumulant LBM) [9], ELBM (Entropic LBM) [1], KBC (Entropic multi-relaxation time LBM) [12] etc. LBM can be massively parallelized [17] and, thus, thanks its efficiency, LBM has been already used in several fluid flow problems, such as fluid-solid interaction problems [4, 3], multiphase, multi-component, and porous media flows [10], chemically reacting flows [16], solution of phase-field equations [5], etc.

In this work, we mainly focus on CuLBM.

4 Benchmarks

In this section, benchmark problems for the turbulent boundary layer flow are investigated. We start with the problem of turbulent fluid flow between two parallel plates, particularly, we study the boundary layer above a smooth surface.

The other experiment investigates the boundary layer above a rough surface and is inspired by the project GAČR. no 19-09539S entitled Large structures in the boundary layers over complex surfaces in high Reynolds numbers.

In all cases, only initialization with equilibrium density distribution function is used.

4.1 Flow above a smooth surface

The geometry of the test case is shown in Fig 2. The dimensions of the computational domain Ω are $H = \pi$ m, $L = 2\pi$ m and $W = 2$ m. Ω has periodic boundary conditions in x and y directions. On the walls parallel to the (x, z) plane, a no-slip condition is prescribed using both the bounce-back condition and the interpolated bounce-back conditions, see [13, 9].

Inside the channel, the fluid is accelerated in the direction x by the force $\vec{g} = (g, 0, 0)^T$, $g = 0.0324\ \mathrm{m\,s^{-2}}$. The value of the acceleration is chosen to satisfy $Re_\tau = 180$ with $\nu = 0.001\ \mathrm{m^2\,s^{-1}}$. Initially, the turbulent velocity field provided by CLMM was prescribed. For the experiment with the bounce-back condition, we set mesh parameter $h \doteq 1 \cdot 10^{-2}$ m, and for the experiment with interpolated bounce back condition, we set $h \approx 13 \cdot 10^{-3}$ m. The results are compared with the results from [11].

Based on the choice of the boundary condition for the wall interaction, the results shown in Figs. 3 and 4 were obtained.

For the smooth wall, LBM produces comparable results to the results given in [11]. Nevertheless, in Fig. 3, the magnitude of u^+ is lower than in [11]. This can be caused by the time averaging, because only small time interval was chosen in our simulations.

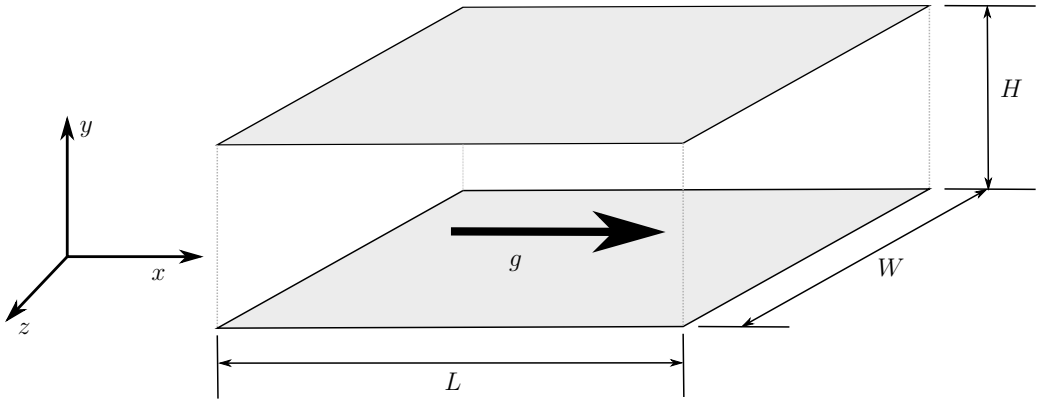


Figure 2: Geometry of the computational domain Ω for the problem of the flow above a smooth surface.

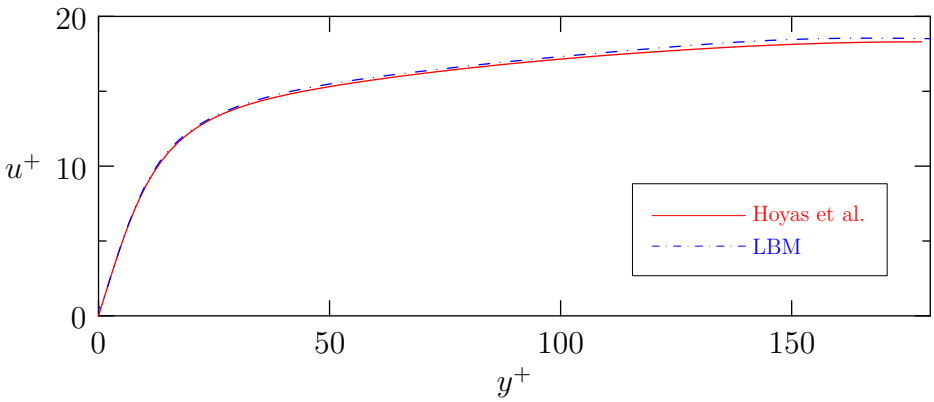


Figure 3: Comparison of the results obtained by LBM with the bounce-back boundary condition and results in [11].

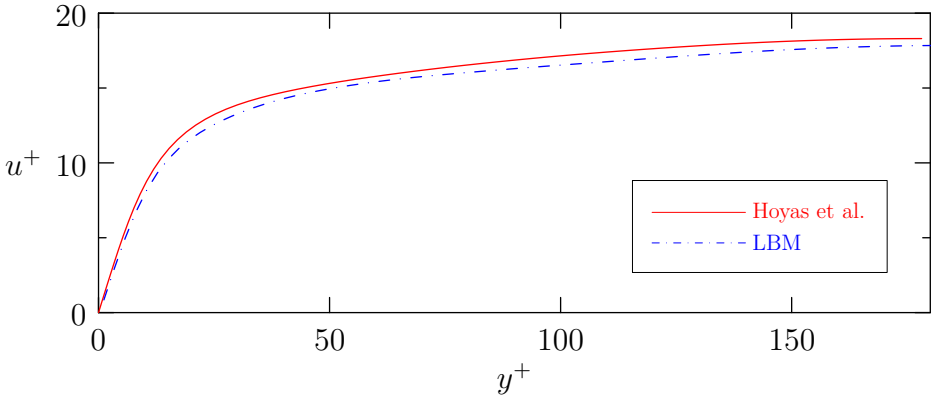


Figure 4: Comparison of results reach by LBM with the interpolated bounce-back boundary condition and results in [11].

4.2 Flow above a rough surface

The computational domains Ω_1 and Ω_2 are shown in Figs. 5a and 5b, respectively. The dimensions (in m) of the computational domain Ω_1 are $(0, 0.207) \times (0, 0.05) \times (0, 0.084)$.

The dimensions (in m) of the computational domain Ω_2 are $(0, 0.514) \times (0, 0.1) \times (0, 0.168)$. Small protrusions are placed at the bottom wall as shown in Figs 5a and 5b. The dimensions (in m) of each protrusion are $0.001 \times 0.004 \times 0.004$ in x, y, and z direction, respectively. Ω_1 and Ω_2 have periodic boundary conditions in the x and z directions. At the walls, the no-slip boundary condition is prescribed using the bounce-back boundary condition. At the boundary parallel to the (x, z) plane, $y = 0.05$ m for Ω_1 and $y = 0.1$ m for Ω_2, symmetric boundary condition is prescribed. The initial condition is $\vec{u}(\vec{x}) = \vec{0}\ \mathrm{m\,s^{-1}}$ and $\rho(\vec{x}) = 1\ \mathrm{kg\,m^{-3}}$ for $\forall \vec{x} \in \Omega_k$, $k = 1, 2$. The mesh parameter $h = 2 \cdot 10^{-4}$ m for Ω_1 and $h = 4 \cdot 10^{-4}$ m for Ω_2.

The fluid is accelerated by the volume force $\vec{g} = (0.2, 0, 0)^T\, m\, s^{-2}$, the kinematic viscosity is $\nu = 1/70000\ \mathrm{m^2\,s^{-1}}$, $u_\tau = 0.1\ \mathrm{m\,s^{-1}}$, and $Re_\tau = 350$.

Fig. 6 illustrates the results of CLMM and LBM in the computational domain Ω_1. There are small differences in the velocity magnitude. The differences can be caused by the averaging procedure because in CLMM, the quantity $\overline{u}_x$ was averaged over 2 s while in LBM, the average was computed over 4 s with the initial time of $t = 40$ s.

Fig. 7 shows the results of CLMM and LBM in the computational domain Ω_2. It can be seen that as y grows, the differences between CLMM and LBM grows. But if we rescale the results by LBM in order to have the same value in $y = 0.1$ m as CLMM, the differences are not so significant, see Fig. 8. The differences can be caused by different averaging procedures because in CLMM, the quantity $\overline{u}_x$ was averaged over 2 s and in LBM over 4 s with the initial time of $t = 100$ s. Because single precision was used for numbers computer arithmetic, the results in Figs. 6, 7, and 8 which are averaged both in space and time can suffer from numerical errors caused by lower precision. Finally, different initial condition were used in our LBM code. Nevertheless, we can state that LBM provides good results in comparison with CLMM code.

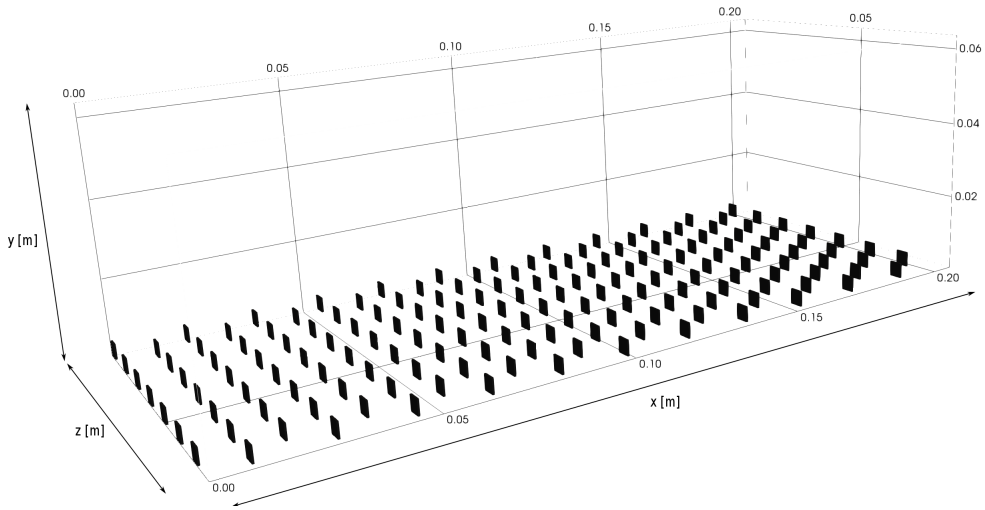
5 Conclusion and future work

We presented a computational study of turbulent fluid flow in the boundary layer using LBM. The primary aim of this work was to introduce the boundary layer theory, LBM, various boundary and initial conditions in the mesoscopic description, and applications on benchmark problems.

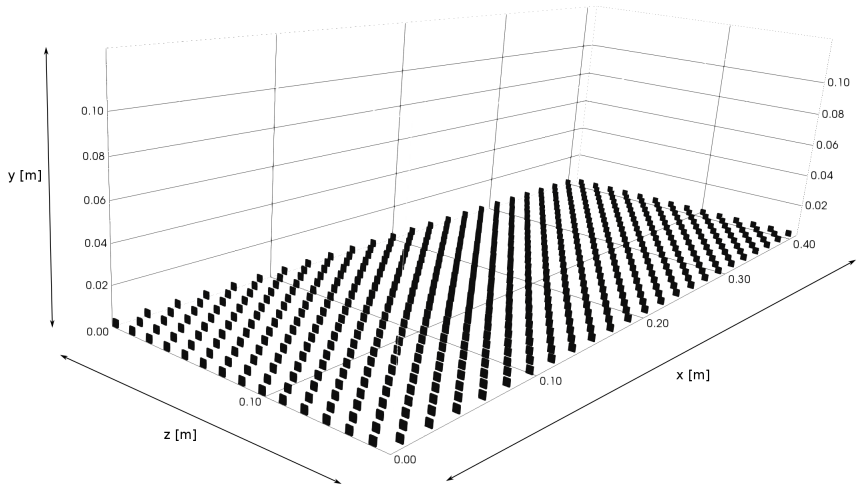
We used two benchmark tests. The first benchmark problem, inspired by [11], represents the fluid flow interaction with a smooth wall. The vertical profile of u^+ was investigated. The profile near the wall matched the experimental data from [11]. As the distance from the wall increased, the differences between the experiment and LBM results increased too.

The second experiment was inspired by the GAČR project. The obtained numerical results are in a good agreement with the data produced by the code CLMM.

We observed that LBM produces satisfactory results in comparison with literature or other solvers in both experiments. Nevertheless, the results differ as the distance from the wall grows. This observation can be caused by several aspects. The first aspect can be the insufficient resolution of the computational mesh. Next, the results were time-averaged over different periods of time, which can influence the results too.



(a) The geometry of the computational domain Ω_1 for the problem of flow above a rough surface with protrusions of dimensions (in m) $0.001 \times 0.004 \times 0.004$ in x, y, and z direction, respectively.



(b) The geometry of the computational domain Ω_2 for the problem of flow above a rough surface with protrusions of dimensions (in m) $0.001 \times 0.004 \times 0.004$ in x, y, and z direction, respectively.

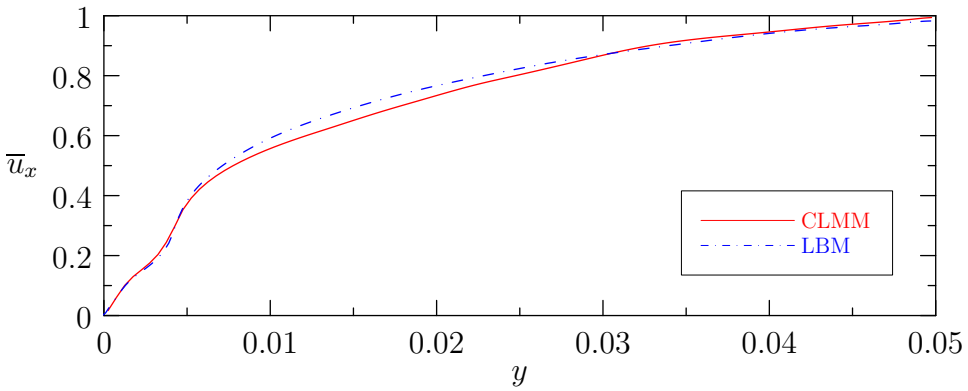


Figure 6: A comparison of results for time-averaged $\overline{u}_x$ obtained by LBM with the bounce-back boundary condition and results provided by the CLMM code without rescaling of the results in the computational domain Ω_1.

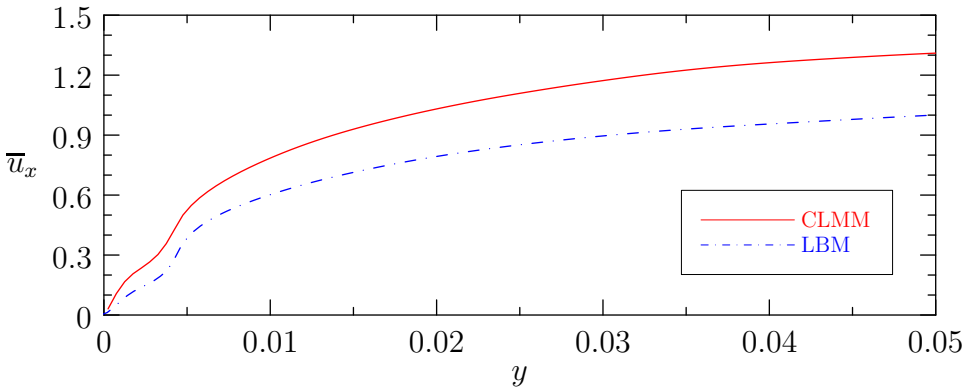


Figure 7: A comparison of results for time-averaged $\overline{u}_x$ obtained by LBM with the bounce-back boundary condition and results provided by the CLMM code without rescaling of the results in the computational domain Ω_2.

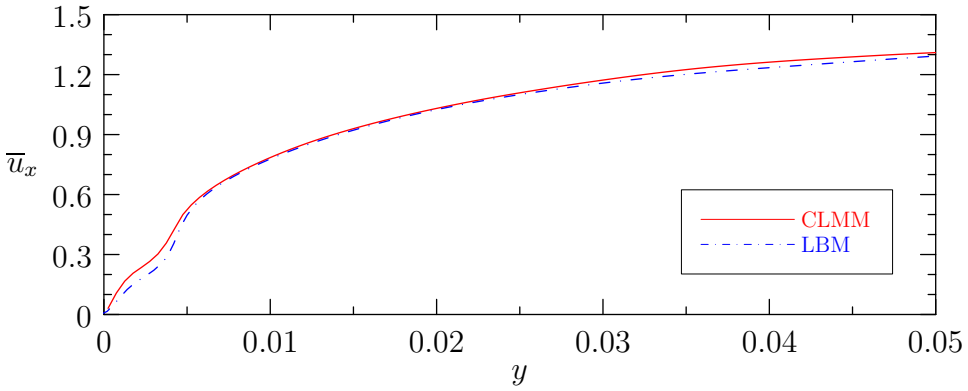


Figure 8: A comparison of rescaled results for time-averaged $\overline{u}_x$ obtained by LBM with the bounce-back boundary condition and results provided by the CLMM code without rescaling of the results in the computational domain Ω_2.

References

[1] S. Chikatamarla, S. Ansumali, and I. V. Karlin. *Entropic lattice Boltzmann models for hydrodynamics in three dimensions.* Physical review letters **97** (2006), 010201.

[2] D. d'Humieres. *Generalized lattice-Boltzmann equations.* Rarefied gas dynamics (1992).

[3] P. Eichler, R. Fučík, and R. Straka. *Computational study of immersed boundary – lattice Boltzmann method for fluid-structure interaction.* Under review in DCDS-S .

[4] R. Fučík, P. Eichler, R. Straka, P. Pauš, J. Klinkovský, and T. Oberhuber. *On optimal node spacing for immersed boundary–lattice Boltzmann method in 2D and 3D.* Computers & Mathematics with Applications **77** (2019), 1144–1162.

[5] M. Geier, A. Fakhari, and T. Lee. *Conservative phase-field lattice Boltzmann model for interface tracking equation.* Physical Review E **91** (2015), 063309.

[6] M. Geier, A. Greiner, and J. G. Korvink. *Cascaded digital lattice Boltzmann automata for high Reynolds number flow.* Physical Review E **73** (2006), 066705.

[7] M. Geier, A. Greiner, and J. G. Korvink. *A factorized central moment lattice Boltzmann method.* The European Physical Journal Special Topics **171** (2009), 55–61.

[8] M. Geier, A. Pasquali, and M. Schönherr. *Parametrization of the cumulant lattice Boltzmann method for fourth order accurate diffusion Part II: Application to flow around a sphere at drag crisis.* Journal of Computational Physics **348** (2017), 889–898.

[9] M. Geier, M. Schönherr, A. Pasquali, and M. Krafczyk. *The cumulant lattice Boltzmann equation in three dimensions: Theory and validation.* Computers & Mathematics with Applications **70** (2015), 507–547.

[10] Z. Guo and C. Shu. *Lattice Boltzmann method and its applications in engineering*, volume 3. World Scientific, (2013).

[11] S. Hoyas and J. Jiménez. *Reynolds number effects on the Reynolds-stress budgets in turbulent channels.* Physics of Fluids **20** (2008), 101511.

[12] I. V. Karlin, F. Bösch, and S. Chikatamarla. *Gibbs' principle for the lattice-kinetic theory of fluid dynamics.* Physical Review E **90** (2014), 031302.

[13] T. Krüger, H. Kusumaatmaja, A. Kuzmin, O. Shardt, G. Silva, and E. M. Viggen. *The lattice Boltzmann method.* Springer International Publishing **10** (2017), 978–3.

[14] H. Schlichting and K. Gersten. *Boundary-layer theory.* Springer, (2016).

[15] K. V. Sharma, R. Straka, and F. W. Tavares. *New Cascaded Thermal Lattice Boltzmann Method for simulations of advection-diffusion and convective heat transfer.* International Journal of Thermal Sciences **118** (2017), 259–277.

[16] S. Succi. *The lattice Boltzmann equation: for fluid dynamics and beyond.* Oxford university press, (2001).

[17] J. Tölke. *Implementation of a Lattice Boltzmann kernel using the Compute Unified Device Architecture developed by nVIDIA.* Computing and Visualization in Science **13** (2010), 29.

Predictions of Average Incomes in Small Areas

Ondřej Faltys

4th 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. The empirical best predictor (EBP) under an area–level gamma mixed model is derived and a corresponding algorithm for calculation of parameters of interest is supplied. The model is applied to Spanish Living Condition Survey (SLCS) data from the year 2008 for estimation of averages incomes in small areas. Subsequently, the data are fitted by the known Fay–Herriot model and the quality of EBPs of the parameters of interest under both models is compared. The behaviour of both models seems to be very similar.

Keywords: Empirical best predictor, gamma distribution, parametric bootstrap.

Abstrakt. Pro zobecněný lineární smíšený model na úrovni oblastí mající odezvy z gamma rozdělení odvozujeme empiricky nejlepší prediktor zkoumaných parametrů a poskytujeme algoritmus pro jejich výpočet. Model následně aplikujeme na data ze španělských provincií týkající se životních podmínek z roku 2008, abychom predikovali průměrnou mzdu v malých oblastech. Za stejným účelem pak na zmíněná data aplikujeme známý Fay–Herriotův model a porovnáváme kvalitu empiricky nejlepších predikcí obou modelů. Získané predikce jsou velice podobné pro oba modely.

Klíčová slova: Empiricky nejlepší prediktor, gamma rozdělení, parametrický bootstrap.

1 Introduction

Small area estimation (SAE) is a field of statistic that is closely related to sampling–survey. The beginning of this discipline can be dated to the $11th$ century England and to the $17th$ century Canada, but the major development of SAE–methods has begun in the last decade of the $20th$ century and continues until these days. The basic task of SAE can be formulated as providing reliable estimates of characteristics of interest for so–called *small areas*.

Under an *area*, we can imagine either a real geographic area as a state, province, county, district, etc., or a socio–demographic group that is usually defined by crossing factors as age, sex, working status, education, permanent residence, etc. The word *small* reflects the fact that we do not have enough amount of data to provide a reliable *direct* estimate of a characteristic of interest for the given area. Under the direct estimate, we mean such estimate that uses values of interest only from that area, i.e. not from other areas. On the contrary, there is an *indirect* estimate often employed in SAE that uses values also from other areas. An example of a characteristic of interest can typically be a total or a mean, but it can also be a non–linear function of variables of interest.

SAE models are classified into two categories depending on the level of aggregation of the data. If auxiliary data are available for each sampled individual, then we can use an *unit–level* model that relates the sampled values y_{dj}, $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, to the corresponding auxiliary data. If auxiliary data are not available at the unit–level and we have got only the responses y_{dj}, then we compute a direct estimate per area and model these estimates by auxiliary data expressing proportional representations of some quantities, such as age, sex, education, etc. for each area. As an example, we can consider gender, if there is 60% of women and 40% of men in one particular area, then the corresponding values of regressors are 0.6 and 0.4 for the given area. This kind of SAE models is called *area–level* models. A comprehensive introduction into SAE models can be found e.g. in [7] or [6].

2 Fay–Herriot model

The basic area–level model was developed in [3] for estimation of incomes in small areas with population less than 1000 in the United States. This model has got a fundamental importance between area–level models and various modifications of it were investigated.

Let $\overline{Y}_d$ be a quantity of interest that is to be predicted under the Fay–Herriot model, i.e.

$$\mu_d = \overline{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj} \,, \quad d = 1, \ldots, D \,, \tag{1}$$

where y_{dj} is e.g. an income of the jth person in the area U_d. By sampling data from the population, there is $n_d < N_d$ observations available per area. Let y_d denote a direct estimate of μ_d, i.e.

$$y_d = \frac{1}{\widehat{N}_d} \sum_{j=1}^{n_d} w_{dj} y_{dj} \,, \quad d = 1, \ldots, D \,, \tag{2}$$

where $\widehat{N}_d = \sum_{j=1}^{n_d} w_{dj}$ and w_{dj} is the sampling weight for the jth individual from the area U_d. The Fay–Herriot model has got two levels. The first level is called a *sampling* model and takes the form

$$y_d = \mu_d + e_d \,, \quad d = 1, \ldots, D \,, \tag{3}$$

where $e_d \sim N(0, \sigma_d^2)$ are independent sampling errors with known variances σ_d^2. In practice, σ_d^2 are design–based variances of the direct estimates y_d, $d = 1, \ldots, D$. The second level model links μ_d to auxiliary data as

$$\mu_d = \boldsymbol{x}_d^\top \boldsymbol{\beta} + v_d \,, \quad d = 1, \ldots, D \,, \tag{4}$$

where $\boldsymbol{x}_d^\top = (x_{d1}, \ldots, x_{dp})$ is a row vector of p auxiliary variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a vector of regression parameters and $v_d \sim N(0, \sigma_v^2)$ are identically independently distributed random effects with an unknown variance σ_v^2. Model (4) is called a *linking* model and makes sense provided that y_d are independent *normally* distributed, see [2]. By combination of the sampling and linking model, we get the final model

$$y_d = \boldsymbol{x}_d^\top \boldsymbol{\beta} + v_d + e_d \,, \quad d = 1, \ldots, D \,. \tag{5}$$

Model (5) can be recognized as a LME model. The general formula of a LME model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e}\,, \tag{6}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are matrices of known constants, $\boldsymbol{v}$ and $\boldsymbol{e}$ are normally distributed vectors of random effects and random errors, respectively, and $\boldsymbol{y}$ is a normally distributed vector of responses. If we denote $\mathrm{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}$ and $\mathrm{var}(\boldsymbol{v}) = \boldsymbol{\Sigma}_v$, then for the *best linear unbiased prediction* (BLUP) of μ_d, $d = 1, \ldots, D$, we are then able to derive

$$\widehat{\mu}_d^{blup} = \boldsymbol{l}^\top \widehat{\boldsymbol{\beta}} + \boldsymbol{m}^\top \widehat{\boldsymbol{v}}\,, \tag{7}$$

where

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}\,, \quad \widehat{\boldsymbol{v}} = \boldsymbol{\Sigma}_v \boldsymbol{Z}^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\,, \tag{8}$$

and $\boldsymbol{l}$ and $\boldsymbol{m}$ are vectors of known constants.

Finally, for the BLUP of the parameter μ_d, we get

$$\widehat{\overline{Y}}_d^{blup} = \widehat{\mu}_d^{blup} = \boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{v}_d = \boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \frac{\sigma_v^2}{\sigma_v^2 + \sigma_d^2}(y_d - \boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}}) = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_d^2} y_d + \frac{\sigma_d^2}{\sigma_v^2 + \sigma_d^2} \boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}}. \tag{9}$$

However, in practice we do not know the variance parameter σ_v^2 and therefore we must estimate it. The *empirical best linear unbiased prediction* (EBLUP) $\widehat{\overline{Y}}_d^{eblup}$ is then obtained by substituting σ_v^2 by its estimate $\widehat{\sigma}_v^2$.

As a measure of the precision of the EBLUP $\widehat{\overline{Y}}_d^{eblup}$, the mean squared error

$$\mathrm{MSE}\left(\widehat{\overline{Y}}_d^{eblup}\right) = \mathrm{E}\left(\widehat{\overline{Y}}_d^{eblup} - \overline{Y}_d\right)^2 \tag{10}$$

is employed.

3 Area level gamma mixed model

In this chapter, we present some basic results that we derived for an area–level model with conditional responses from gamma distributions. As a motivation for investigation of this model (further referred as the gamma model), we can mention an estimation of average incomes for small areas, i.e. we suppose that every average income for a small area follows a gamma distribution. Another important feature of this model is that the parameters of interest (the average incomes) are supposed to follow asymmetric distributions (particularly the gamma distr.). This is in contrast with the presented Fay–Herriot model which assumes that these parameters are normally distributed and thus from symmetric distributions.

3.1 Gamma model

Let us consider the random effects $v_1, \ldots, v_D \underset{iid}{\sim} N(0, 1)$, then the multivariate density of a vector $\boldsymbol{v} = (v_1, \ldots, v_D)^\top$ is

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

By these effects, we try to capture the variability of responses that is not explained by auxiliary data. Next, let the distribution of the target variable y_d, conditioned to the random effect v_d, be

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

where ν_d is the *shape* parameter of the gamma distribution and it is assumed to be known. We also assume that y_d are independent random variables for $d = 1, \ldots, D$. For the expectation and variance of y_d given v_d, we have got

$$E(y_d|v_d) = \frac{\nu_d}{a_d} = \mu_d, \quad \text{var}(y_d|v_d) = \frac{\nu_d}{a_d^2} = \frac{\mu_d^2}{\nu_d}\,,$$

and the density takes the form

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

Since the canonical link for the gamma distribution is the inverse link $g(\mu) = \frac{1}{\mu}$, then we get the following GLME model

$$g(\mu_d) = \frac{1}{\mu_d} = \boldsymbol{x}_d^\top \boldsymbol{\beta} + \phi v_d, \quad d = 1, \ldots, D, \tag{11}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a vector of regression parameters and $\boldsymbol{x}_d^\top = (x_{d1}, \ldots, x_{dp})$ is a row vector of area–level auxiliary data for the area U_d. Model (11) is called the *area–level gamma mixed model* (abbreviated as the gamma model).

Let $\boldsymbol{\theta}$ denote the unknown parameters of model (11), i.e. $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi)^\top$. Assuming that $y_d|v_d$ are independent, i.e. $f(\boldsymbol{y}|\boldsymbol{v}) = \prod_{d=1}^D f(y_d|v_d)$, the *likelihood* function $L(\boldsymbol{\theta})$ can be expressed as

$$L(\boldsymbol{\theta}) = f(\boldsymbol{y}) = \int_{\mathbb{R}^D} f(\boldsymbol{y}|\boldsymbol{v}) f_{\boldsymbol{v}}(\boldsymbol{v}) \mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^D} \psi(\boldsymbol{y}, \boldsymbol{v}) \mathrm{d}\boldsymbol{v}, \tag{12}$$

where

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

There are $p + 1$ unknown parameters in model (11): the regression parameters, $\beta_1, \ldots, \beta_p$, and ϕ. Since we are not able to calculate the integral in (12) explicitly, we cannot employ the classical *maximum likelihood estimation* (MLE) method for parameter estimation. Instead, we approximate the integral and calculate MLEs of the unknown parameters for this approximation.

3.2 Empirical best predictor

Providing reliable predictions of the parameters of interest, μ_d, $d = 1, \dots, D$, under gamma model (11) is of crucial importance. The *best predictor* (BP) $\widehat{\mu}_d^{BP}$ of $\mu_d = \mu_d(\boldsymbol{\theta}, v_d)$ minimizes the mean squared error in the class of predictors $\widehat{\mu} = \widehat{\mu}(s)$ depending on the sample s, i.e.

$$\widehat{\mu}^{BP} \in \operatorname*{argmin}_{\widehat{\mu}} \mathrm{E}(\widehat{\mu} - \mu)^2 \,.$$

Besides of the sample s, $\widehat{\mu}_d^{BP}$ depends on $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi)^\top$. It can be shown that $\widehat{\mu}_d^{BP}(\boldsymbol{\theta}) = E_\theta(\mu_d|\boldsymbol{y})$, see [5]. Moreover, it obviously holds $E_\theta(\mu_d|\boldsymbol{y}) = E_\theta(\mu_d|y_d)$ under model (11). Thus, we get

$$\widehat{\mu}_d^{BP}(\boldsymbol{\theta}) = E_\theta(\mu_d|y_d)\,, \quad d = 1, \dots, D. \tag{13}$$

The right side of (13) can be expressed as

$$E_\theta[\mu_d|y_d] = \frac{\int_R (\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d)^{-1} f(y_d|v_d) f(v_d) \mathrm{d}v_d}{\int_R f(y_d|v_d) f(v_d) \mathrm{d}v_d} = \frac{\mathcal{N}_d(y_d, \boldsymbol{\theta})}{\mathcal{D}_d(y_d, \boldsymbol{\theta})} = \frac{N_d(y_d, \boldsymbol{\theta})}{D_d(y_d, \boldsymbol{\theta})}\,,$$

where $\mathcal{N}_d = \mathcal{N}_d(y_d, \boldsymbol{\theta})$, $\mathcal{D}_d = \mathcal{D}_d(y_d, \boldsymbol{\theta})$, $N_d = N_d(y_d, \boldsymbol{\theta})$ and $D_d = D_d(y_d, \boldsymbol{\theta})$ are

$$\mathcal{N}_d = \int_R (\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d)^{-1} \exp\left\{\log \frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{\Gamma(\nu_d)} + \nu_d \log(\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d) - \nu_d y_d (\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d)\right\} f(v_d)\mathrm{d}v_d\,,$$

$$\mathcal{D}_d = \int_R \exp\left\{\log \frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{\Gamma(\nu_d)} + \nu_d \log(\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d) - \nu_d y_d (\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d)\right\} f(v_d)\mathrm{d}v_d\,,$$

$$N_d = \int_R (\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d)^{-1} \exp\left\{\nu_d \log(\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d) - \nu_d y_d (\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d)\right\} f(v_d)\mathrm{d}v_d\,,$$

$$D_d = \int_R \exp\left\{\nu_d \log(\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d) - \nu_d y_d (\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d)\right\} f(v_d)\mathrm{d}v_d\,.$$

All these integrals have to be approximated numerically. The best predictor of v_d can be also derived in the same way as

$$\widehat{v}_d^{BP}(\boldsymbol{\theta}) = E_\theta[v_d|y_d] = \frac{\int_R v_d f(y_d|v_d) f(v_d) \mathrm{d}v_d}{\int_R f(y_d|v_d) f(v_d) \mathrm{d}v_d} = \frac{N_{v,d}(y_d, \boldsymbol{\theta})}{D_d(y_d, \boldsymbol{\theta})}\,,$$

where

$$N_{v,d}(y_d, \boldsymbol{\theta}) = \int_R v_d \exp\left\{\nu_d \log(\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d) - \nu_d y_d (\boldsymbol{x}_d^\top\boldsymbol{\beta} + \phi v_d)\right\} f(v_d)\mathrm{d}v_d\,.$$

The *empirical best predictor* (EBP) of μ_d is then obtained by substituting $\boldsymbol{\theta}$ for its estimate $\widehat{\boldsymbol{\theta}}$ in (13) and we denote it as $\widehat{\mu}_d^{EBP} \equiv \widehat{\mu}_d^{BP}(\widehat{\boldsymbol{\theta}})$.

3.2.1 Algorithm

Integrals given above, $\mathcal{N}_d$, $\mathcal{D}_d$, N_d and D_d, cannot be calculated explicitly. Thus, they must be approximated numerically by Monte Carlo methods. The following procedure provides the EBP of μ_d and v_d for $d = 1, \dots, D$.

1. Estimate $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\phi})^\top$.
2. For $s = 1, \ldots, S$: generate $v_d^{(s)} \underset{iid}{\sim} N(0,1)$ and put $v_d^{(S+s)} = -v_d^{(s)}$.
3. Calculate the EBP $\widehat{\mu}_d^{BP}(\widehat{\boldsymbol{\theta}}) = \widehat{N}_d / \widehat{D}_d$ where

$$\widehat{N}_d = \frac{1}{2S} \sum_{s=1}^{2S} (\boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{\phi} v_d^{(s)})^{-1} \exp\left\{ \nu_d \log(\boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{\phi} v_d^{(s)}) - \nu_d y_d (\boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{\phi} v_d^{(s)}) \right\},$$

$$\widehat{D}_d = \frac{1}{2S} \sum_{s=1}^{2S} \exp\left\{ \nu_d \log(\boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{\phi} v_d^{(s)}) - \nu_d y_d (\boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{\phi} v_d^{(s)}) \right\}.$$

4. Calculate the EBP $\widehat{v}_d^{BP}(\widehat{\boldsymbol{\theta}}) = \widehat{N}_{v,d} / \widehat{D}_d$, where

$$\widehat{N}_{v,d} = \frac{1}{2S} \sum_{s=1}^{2S} v_d^{(s)} \exp\{ \nu_d \log(\boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{\phi} v_d^{(s)}) - \nu_d y_d (\boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{\phi} v_d^{(s)}) \}.$$

The calculation of $\widehat{\mu}_d^{EBP}$, $d = 1, \ldots, D$, is computationally demanding. There is a simple predictor of $\mu_d = \mu_d(\boldsymbol{\theta}, v_d)$ that is obtained by substituting $\boldsymbol{\theta}$ and v_d by parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\phi})^\top$ and a random effect predictor $\widehat{v}_d$, i.e.

$$\tilde{\mu}_d = (\boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{\phi} \widehat{v}_d)^{-1}, \quad d = 1, \ldots, D.$$

It is called *plug-in* predictor.

3.2.2 Bootstrap estimation of MSE

After obtaining the EBP of μ_d ($d = 1, \ldots, D$) under the gamma model, it is necessary to assess its precision. Particularly, we are interested in derivation of the MSE of the predictor $\widehat{\mu}_d^{EBP}$, i.e.

$$\mathrm{MSE}(\widehat{\mu}_d^{EBP}) = \mathrm{E}(\widehat{\mu}_d^{EBP} - \mu_d)^2, \quad d = 1, \ldots, D, \tag{14}$$

where the expectation is taken with respect to gamma model (11). Unfortunately, the explicit expression of (14) does not exist and only an approximation of (14) can be searched for. For this reason, we employ the parametric bootstrap described in [4] which enables us easily to compute estimates of $\mathrm{MSE}(\widehat{\mu}_d^{EBP})$, $d = 1, \ldots, D$. The following procedure calculates these estimates.

1. Fit the model to the sample and calculate the estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\phi})^\top$.
2. Repeat B times ($b = 1, \ldots, B$):
   (a) Generate $v_d^{(b)} \sim N(0,1)$, $y_d \sim \mathrm{Gamma}\left(\nu_d, \frac{\nu_d}{\mu_d^{(b)}}\right)$, $\mu_d^{(b)} = (\boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{\phi} v_d^{(b)})^{-1}$, $d = 1, \ldots, D$.
   (b) For each bootstrap sample, calculate the estimator $\widehat{\boldsymbol{\theta}}^{(b)}$ and the EBP $\widehat{\mu}_d^{EBP\,(b)} = \widehat{\mu}_d^{BP}(\widehat{\boldsymbol{\theta}}^{(b)})$.
3. Output: $mse(\widehat{\mu}_d^{EBP}) = \frac{1}{B} \sum_{b=1}^{B} (\widehat{\mu}_d^{EBP\,(b)} - \mu_d^{(b)})^2$.

4 Application to real data

As suitable data for application of model (11), we choose the Spanish Living Condition Survey (SLCS) data from the year 2008. To create some idea of what data it is, we briefly describe it. The data were obtained from 52 Spanish provinces by conducting a survey. Of all the information found in the survey, we are especially interested in personal incomes. Depending on gender, we compute a direct estimate of the average income for both men and women per province, i.e. we get the estimates for $2 * 52 = 104$ socio–demographic groups (small areas). These values are denoted as y_d, $d = 1, \ldots, D = 104$. Auxiliary data then correspond to relative representations of various socio–demographic characteristics such as sex, education, working status, etc. per area.

We fit the prepared data both by gamma model (11) and also by Fay–Herriot model (5) where the direct estimates are treated as normally distributed. Subsequently, we compare the outputs of both of them. Employed regressors at the area–level are: the proportion of unemployed individuals, of individuals aged from 16 to 24, of individuals aged 65 and older and of individuals with university education per area. These variables are denoted as $x_{d1}, \ldots, x_{d4}$, respectively. The outputs of both models are summarized in Tables 3.13 and 3.14.

Gamma model				
	Estimate	Standard error	Wald (t-test) statistic	p-value
$\widehat{\beta}_0$	1.58e-01	5.14e-02	9.39	2.18e-03
$\widehat{\beta}_1$	1.90	2e-01	91.81	9.53e-22
$\widehat{\beta}_2$	3.28	3.95e-01	68.89	1.04e-16
$\widehat{\beta}_3$	1.43	8.14e-02	306.48	1.28e-68
$\widehat{\beta}_4$	-9.3e-01	8.25e-02	127.14	1.73e-29
$\widehat{\phi}$	5.87e-02	3.15e-03	348.67	8.27e-78

Table 3.13: Regression parameter estimates for gamma model (11).

Fay-Herriot model				
	Estimate	Standard error	Wald (t-test) statistic	p-value
$\widehat{\beta}_0$	2.44	2.49e-01	9.82	9.18e-23
$\widehat{\beta}_1$	-3.71	8.72e-01	-4.25	2.16e-05
$\widehat{\beta}_2$	-6.12	1.79	-3.43	6.11e-04
$\widehat{\beta}_3$	-2.78	3.61e-01	-7.71	1.27e-14
$\widehat{\beta}_4$	2.20	5.08e-01	4.33	1.47e-05
$\widehat{\phi}$	1.19e-01			

Table 3.14: Regression parameter estimates for Fay–Herriot model (5).

From Tables 3.13 and 3.14, we can see that all employed regressors are significant at the 5% significance level for both models. The residuals for the gamma model are computed as

$$r_d = y_d - \widehat{\mu}_d\,, \quad d = 1, \ldots, D\,, \tag{15}$$

where $\widehat{\mu}_d = (\boldsymbol{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{\phi}\widehat{v}_d)^{-1}$ is in fact the plug-in predictor. The residual sum of squares is then

$$r^2 = \sum_{d=1}^{D} r_d^2 = 0.321\,.$$

For the fit by the Fay–Herriot model, we employed the function `eblupFH` from the package sae implemented in the R software for statistical computing. Unfortunately, this function does not deliver predictions of the random effects, $\widehat{v}_d$, $d = 1, \ldots, D$, but only an estimate $\widehat{\sigma_v^2}$ of the variance of these effects. For this reason, we are not able to compute the residuals for the Fay–Herriot model in the same way as for the gamma model, see (15). Thus, we cannot compare the quality of both fits on the basis of the residual sum of squares.

As the next step, we compute the EBP of μ_d, $d = 1, \ldots, D$, under the gamma model, and the EBLUP of μ_d under the Fay–Herriot model by the function `eblupFH`. The results are very similar as evident from Figure 1. The precision of both predictors is depicted in Figure 2, where the estimates of $\mathrm{MSE}(\widehat{\mu}_d)$, $d = 1, \ldots, D$, were calculated by the bootstrap algorithm for the gamma model and by the function `mseFH` from the sae package for the Fay–Herriot model. The function `mseFH` does not use the parametric bootstrap for calculation of the MSE estimates, but under the Fay–Herriot model, an approximation of $\mathrm{MSE}(\widehat{\mu}_d^{EBLUP}) = \mathrm{E}(\widehat{\mu}_d^{EBLUP} - \mu_d)^2$, $d = 1, \ldots, D$, is employed, see [1]. In this figure, it is also plotted the variance of the direct estimates y_d, $d = 1, \ldots, D$.

We can immediately see from Figure 2 that by the introduction of both models, greater precision for small area estimates was achieved compared to the direct estimates y_d. If we compute the means of the MSE estimates depicted in Figure 2 over all domains, we get 3.20e−03 and 3.08e−03 for the Fay–Herriot model and the gamma model, respectively. Thus, the MSE estimates obtained under the gamma model are obviously slightly more accurate than those obtained under the Fay–Herriot model by `mseFH` function.

5 Conclusion

In this work we introduced area–level model (11) assuming that the conditional responses are from gamma distributions. As a suitable motivation for investigation of this model, we can consider e.g. the prediction of average incomes for small areas. Under a *small area*, we mean either a real geographical area or a socio–demographic group for which predictions of characteristics of interest are required and, at the same time, the sample size of observations for this area (or group) is too small to provide the reliable predictions. It means that these predictions usually have got a large variance. For this reason, we define model (11) using data from all areas to improve the direct estimates for each particular area in sense of minimizing the MSE.

As a characteristic of interest, we consider the average income per area. We derive

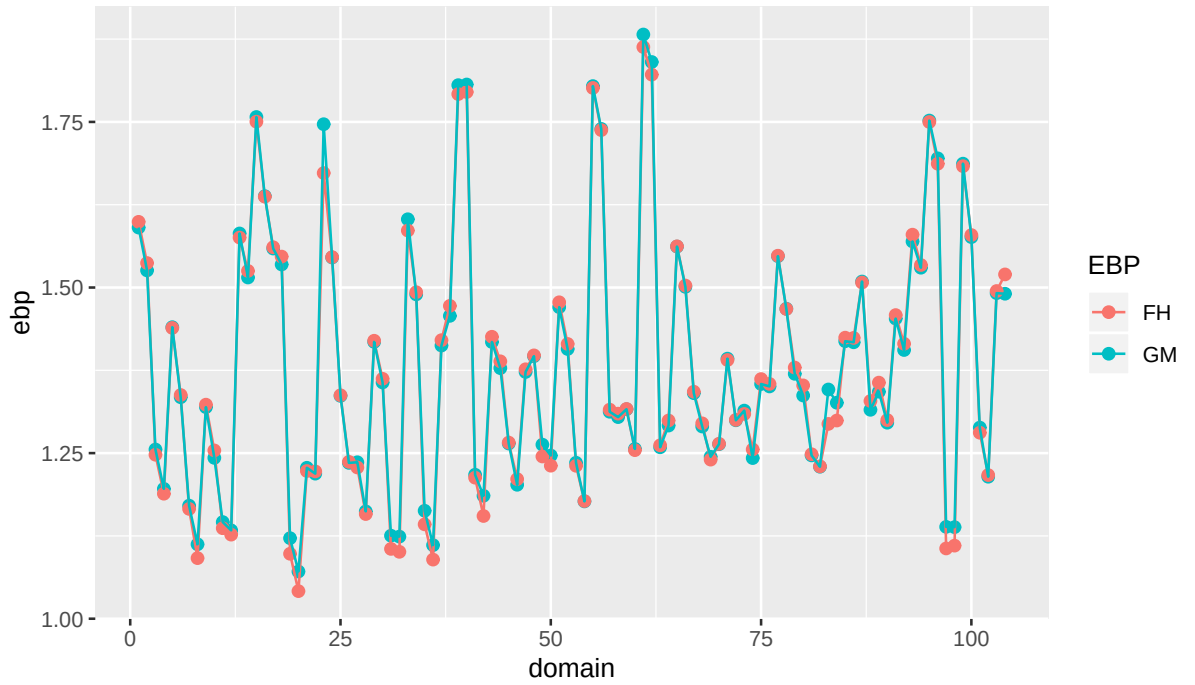


Figure 1: Comparison of the EBPs for the gamma model with the EBLUPs for the Fay–Herriot model. The predictions under both models seem to be very similar.

the *empirical best predictor* (EBP) under this model and provide algorithm 3.2.1 for calculation of the EBPs. Moreover, the plug-in predictor of the average incomes is supplied.

Subsequently, gamma model (11) and known Fay–Herriot model (5) are applied to real data. Specifically, to the Spanish Living Condition Survey (SLCS) data from the year 2008. The summary of both fits is presented in tables 3.13 and 3.14. All used regressors proved to be significant at the 5% significance level for both models as evident from these tables.

The EBPs of the parameters of interest, $\mu_d\,(d = 1, \ldots, D)$, obtained under the gamma model seem to be very similar to the EBLUPs of μ_d under the Fay–Herriot model according to Figure 1. Dealing with the MSE estimates of these predictors, the EBPs may be slightly more accurate than the EBLUPs obtained under the Fay–Herriot model, see Figure 2. Finally, we can conclude that by the introduction of both models, some reduction in the MSE of the direct estimates y_d, $d = 1, \ldots, D$, was achieved as evident from Figure 2 and thus it is reasonable to employ both models for estimation of the average incomes for small areas.

References

[1] G. S. Datta, P. Lahiri. *A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems.* Statistica Sinica (2000), 613–627.

[2] A. J. Dobson. *An introduction to generalized linear models.* Chapman and Hall Statistics Text Series (1990). ISBN 9780412311000.

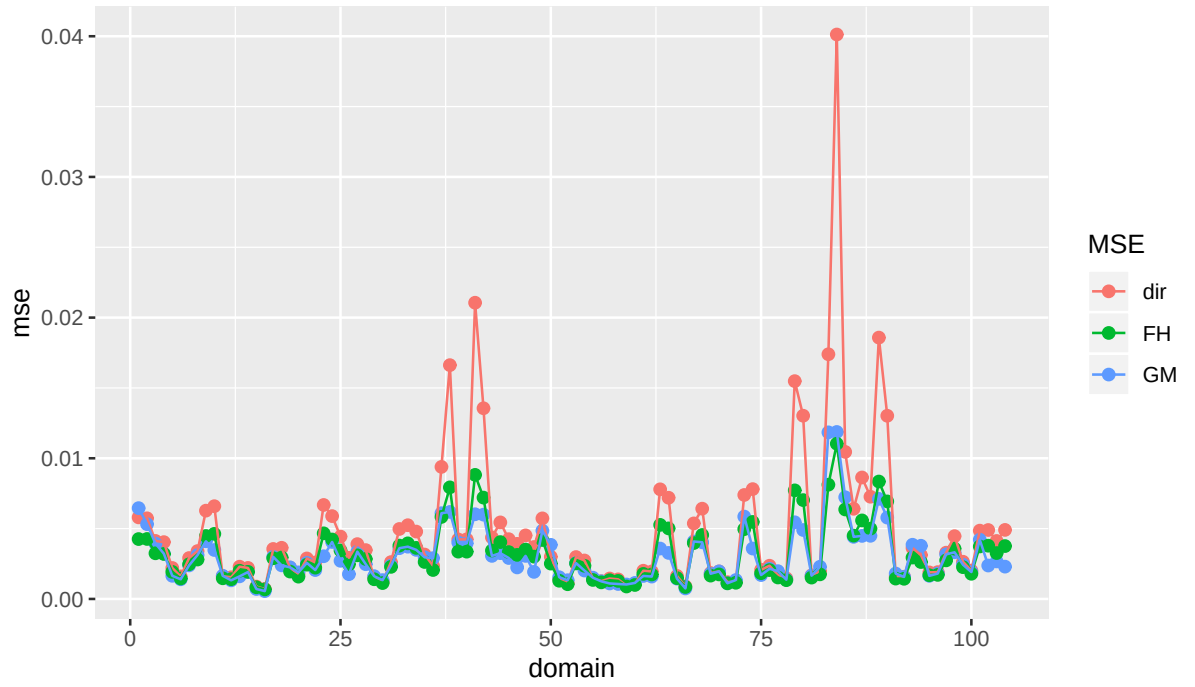


Figure 2: The MSE estimates of the EBPs and the EBLUPs under the gamma model and the Fay–Herriot model, respectively, compared to the variance of the direct estimates.

[3] R. E. Fay, R. A. Herriot. *Estimates of income for small places: an application of james-stein procedures to census data.* Journal of the American Statistical Association **74** (1979), 269–277.

[4] W. González-Manteiga, M. J. Lombardía, I. Molina, D. Morales, L. Santamaría. *Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model.* Computational statistics & data analysis **51** (2007), 2720–2733.

[5] J. Jiang, P. Lahiri. *Empirical best prediction for small area inference with binary data.* Annals of the Institute of Statistical Mathematics **53** (2001), 217–243.

[6] J. N. K. Rao, I. Molina. *Small area estimation, 2nd Edition.* John Wiley, (Hoboken, N.J., 2015), ISBN 978-1-118-73578-7.

[7] J. N. K. Rao. *Small area estimation.* John Wiley, (Hoboken, N.J., 2003), ISBN 0-471-41374-7.

Konstrukce stabilizujících řízení respektující cenu pomocí demonstrací

Jiří Fejlek

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

školitel: Stefan Ratschan, Oddělení výpočetní matematiky
Ústav informatiky, Akademie věd České republiky

Abstract. This paper is concerned with the construction of stabilizing control laws for control-affine systems with feedback using demonstrations and linearization. These demonstrations are trajectories of a stabilized system and are provided via a numerical solver. We construct the feedback control law for the whole state space using these trajectories and test it using Lyapunov functions that are also learned on the demonstrations. Hence the gathered control laws represent stabilizing strategies that may also respect a chosen cost.

Keywords: system stabilization, control-affine systems, Lyapunov functions

Abstrakt. Tento text popisuje konstrukci stabilizujících řízení se zpětnou vazbou pro systémy afinní v řízení, a to pomocí demonstrací a linearizace dynamiky systému. Demonstrace jsou tvořeny trajektoriemi stabilizovaného systému předpočítanými numerickým řešičem. Pomocí těchto trajektorií získáme zpětnovazebné řízení pro stavový prostor. Toto řízení následně ověříme pomocí Lyapunovských funkcí, které též získáme z napočítaných demonstrací. Postup celkově poskytuje stabilizující řízení, která navíc mohou respektovat zvolenou cenu.

Klíčová slova: stabilizace systémů, systémy afinní v řízení, Lyapunovské funkce

1 Úvod

Uvažme systém lineární v řízení označovaný též jako systém afinní v řízení. Standardní úlohou je navrhnout pro takový systém řízení, které stabilizuje tento systém do rovnovážného stavu. Obvyklým postupem v takové úloze je nalezení vhodné Lyapunovy funkce pro systém s řízením (dále zkr. CLF z angl. *control-Lyapunov function*) a pomocí Sontagovy formule [11] získat zpětnovazebné stabilizující řízení.

Častým dodatečným požadavkem na řízení však i je minimalizace ceny podél trajektorie systému. V tu chvíli se však projevuje nevýhoda Sontagovy formule, která obecně negeneruje optimální řízení k předem vybranému kritériu ceny [5]. Proto v tomto textu navrhujeme nový postup generace řízení využívající předpočítané trajektorie získané řešičem založeném na řešení optimálního řízení převodem na úlohu nelineárního programování.

Tento postup je inspirován pracemi [13] a [10], ale navíc dále užijeme napočítané trajektorie pro naučení se CLF dle práce [9]. Toto umožňuje napočítané řízení efektivně kontrolovat a dle potřeby doplňovat vhodně o další demonstrace. Při úspěchu procedury je tedy výsledkem stabilizující řízení, které díky demonstrátoru respektuje i cenu tohoto řízení.

2 Definice problému

Nechť je systém popsaný soustavou diferenciálních rovnic

$$\dot{x} = F(x) + G(x)u, \tag{1}$$

přičemž $x \in \mathbb{R}^n$ označuje stavy systému a $u \in \mathbb{R}^d$ označuje řízení a $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ a $G : \mathbb{R}^n \mapsto \mathbb{R}^{n\times d}$ jsou hladké funkce. O tomto systému navíc předpokládejme, že má v bodě $x = 0$ jediný rovnovážný stav.

Nechť je stav našeho systému různý od rovnovážného a mějme přání dosáhnout pro $t \to +\infty$ rovnovážného stavu. Abychom tak učinili, je obecně nutné působit na systém pomocí *stabilizujícího* řízení $u(t)$. Předpokládejme, že takové řízení pro každý stav systému existuje. Navíc chceme, aby toto řízení bylo zpětnovazebné, tedy tvaru

$$u(t) = k(x(t)), \tag{2}$$

které dále respektuje kvadratický funkcionál ceny

$$V(x(0)) = \int_0^{+\infty} x(t)^T P x(t) + u(t)^T Q u(t)\ \mathrm{d}t, \tag{3}$$

kde P a Q jsou pozitivně definitní matice. Postačující podmínkou pro existenci stabilizujícího řízení je existence spojitě diferencovatelné CLF L splňující pro všechny stavy systému x podmínky [12]

1. $L(x) \to +\infty$ pro $\|x\| \to +\infty$
2. $L(x) \geq 0$ pro všechna x a $L(x) = 0$ právě tehdy, když $x = 0$
3. $\min_u \nabla L(x)^T (F(x) + G(x)u) < 0$ pro všechna $x \neq 0$

Pro systém afinní v řízení navíc lze stabilizující řízení zvolit ve zpětnovazebné podobě explicitně pomocí Sontagovy formule jako [11]

$$u = -G(x)^T \nabla L(x) \frac{\nabla L(x)^T F(x) + \sqrt{(\nabla L(x)^T F(x))^2 + (\nabla L(x)^T G(x) G(x)^T \nabla L(x))^2}}{\nabla L(x)^T G(x) G(x)^T \nabla L(x)}, \tag{4}$$

pokud jmenovatel není nula. Pro tento případ se klade $u = 0$. Závěrem poznamenejme, že stabilizovatelnost nelineárních systémů nás může zajímat pouze na nějaké kompaktní souvislé podmnožině S obsahující počátek. V tomto případě jsou podmínky na CLF analogické globálnímu případu s omezením na množinu S. Pak tato CLF implikuje existenci řízení na jisté podmnožině množiny S, které stabilizuje systém a tuto podmnožinu nikdy neopustí [9]. Pro samotné příklady tedy stačí hledat CLF na „dostatečně“ velké množině stavů obsahující počátek.

3 Koncept konstrukce stabilizujícího řízení

Pro začátek uveďme stručný přehled postupu pro nalezení stabilizujícího řízení. Předpokládejme, že máme na počátku množinu navzorkovaných trajektorií $\mathcal{T}$ představující

stabilizaci našeho systému (1) z vybraných počátečních stavů systému. Tyto trajektorie získáme pomocí *demonstrátoru* založeném např. na numerické optimalizaci.

V první fázi algoritmu se pomocí těchto předpočítaných trajektorií naučíme CLF pro náš systém (1). Následně definujeme s využitím napočítaných trajektorií řízení pro všechny stavy systému užitím linearizace dynamiky systému. Toto řízení následně zkontrolujeme prostřednictvím získané CLF. Pokud toto ověření selže, požádáme demonstrátor o další trajektorie začínající stavy systému, kde byl problém detekován. Tyto trajektorie (a nové demonstrace) využijeme pro aktualizaci CLF a našeho řízení a celý proces opakujeme dokud nenastane úspěch při ověřování. Bude-li celý postup úspěšný, získáme nakonec řízení, které je kompatibilní s CLF pro náš systém, a tedy je stabilizující.

Koncept algoritmu

1. Počáteční systém trajektorií $\mathcal{T}$ obsahující demonstrace $(\mathcal{X},\mathcal{U})$ získaných demonstrátorem.
2. Do nalezení stabilizujícího řízení nebo selhání procedury:
   (a) Nauč se CLF L na vzorcích $(\mathcal{X},\mathcal{U})$, v průběhu učení doplňuj do $\mathcal{T}$ a $(\mathcal{X},\mathcal{U})$ případné nalezené protipříklady s využitím demonstrátoru.
   (b) Na aktuálním $\mathcal{T}$ sestroj řízení u pomocí linearizace.
   (c) Otestuj, že je získané u stabilizující pomocí CLF L. Pokud nalezen protipříklad x, získej z demonstrátoru trajektorii z x a přidej ji do $\mathcal{T}$. Do $(\mathcal{X},\mathcal{U})$ doplň příslušné demonstrace. Následně opakuj celý postup (a)–(c) znovu.

4 Demonstrátor

Fundamentálním stavebním kamenem navrženého postupu bude učení se řízení z demonstrací poskytnutých demonstrátorem. Ideální demonstrátor představuje černou skříňku, která pro zvolený počáteční stav poskytuje stabilizující řízení s přihlédnutím k cenovému funkcionálu (1). Nejsnadněji uchopitelný a v praxi implementovatelný demonstrátor je tvořen řešičem pro optimální řízení založeném na *přímém* postupu [3]. Ten spočívá v řešení optimalizační úlohy nelineárního programování (NLP) tvaru

$$\min f(x)$$

$$\text{za podmínek}$$

$$g(x) = 0.$$

Tuto úlohy získáme z našeho problému diskretizací hledané optimální trajektorie systému $(x^*(t), u^*(t))$ na tvar (x_k^*, u_k^*) (přičemž budeme dále aproximovat nekonečný časový horizont dostatečně dlouhým konečným časovým horizontem). Cílovou funkci NLP úlohy zvolíme jako diskrétní aproximaci funkcionálu ceny (3) (předpokládejme pevnou délku časového kroku, označenou jako h):

$$\frac{1}{2}\left(x_1^T P x_1 + u_1^T Q u_1 + x_N^T P x_N + u_N^T Q u_N\right) + \sum_{k=2}^{N-1} x_k^T P x_k + u_k^T Q u_k. \tag{5}$$

Vazby v úloze nelineárního programování pak vyjadřují vztahy mezi (x_k, u_k) danými zvolenou diskretizací dynamiky systému [3]. Řešením optimalizační úlohy je pak posloupnost (x_k^*, u_k^*), ze kterého případně získáme řízení ve spojitém čase interpolací.

Důležitým faktorem je, že toto řízení je tzv. *feed-forward*, tedy neobsahuje zpětnou vazbu. Pro další úvahy předpokládejme, že pro zvolenou úlohu máme demonstrátor založený na přímém postupu, který generuje stabilizující řízení.

5 Učení se CLF z demonstrací

Cílem této části je najít CLF, pomocí které lze testovat, že je řízení konstruované dále v textu stabilizující pro systém (1). Na rozdíl od prací [13] (využívající přímý výpočet CLF pomocí náročné *sum-of-squares* optimalizace) či [10] (využívající čistě simulace) uvážíme postup představený v [9], který odpovídá naší představě metody postavené na demonstrátoru.

Mějme náš demonstrátor generující stabilizující řízení a množinu demonstrací. Naším cílem je se z množiny demonstrací naučit CLF pro systém (1). Provedeme to uvážením lineárně parametrizovaného systému funkcí $\{L(x,p)\}$ a nalezneme hodnotu parametrů p tak, aby ve všech bodech daných demonstracemi byly splněny podmínky

1. $L(x,p) \to +\infty$ pro $\|x\| \to +\infty$ a $L(0,p) = 0$.

2. $L(x,p) > 0$ pro všechny demonstrace (x,u) nenulové.

3. $\nabla L(x)^T(F(x) + G(x)u) < 0$ pro všechny demonstrace (x,u) nenulové.

Budeme předpokládat, že bod jedna je splněn vhodnou volbou uvažovaných funkcí. Body dva a tři pro lineárně parametrizovaný systému představují soustavu lineárních nerovnic. Ty vytínají v prostoru možných hodnot parametrů mnohostěn (uvažujme prostor možných hodnot parametrů tvaru $p_{\min} \leq p \leq p_{\max}$). Pro výběr jedné konkrétní hodnoty parametrů zvolíme střed vepsaného elipsoidu o maximálním objemu [9], neboť se mimo jiné jedná o konvexní optimalizační problém [4].

Nalezením hodnoty parametrů určíme kandidáta na CLF *kompatibilního* s demonstracemi. Zbývá ověřit, že se jedná skutečně o CLF pro náš systém (1). Zde se spokojíme pro jednoduchost s falzifikátorem, tedy se pokusíme nalézt protipříklad. Falzifikace podmínky v bodu dva je v jádru jednoduchá, nalezneme minimum $L(x,p)$ a ověříme, že není menší než nula. Pro falzifikaci podmínky 3 využijeme toho, že Sontagova formule (4) poskytuje stabilizující řízení, které je kompatibilní s CLF, která řízení vygenerovala. Tudíž je nutně pro řízení u_S vygenerované Sontagovou formulí $\nabla L(x,p)^T(F(x) + G(x)u_S(x)) < 0$ pro všechna $x \neq 0$, což můžeme opět vyšetřit optimalizací.

Nalezneme-li v průběhu falzifikace protipříklad, vezmeme tento protipříklad jako počáteční stav nové trajektorie, kterou vygenerujeme demonstrátorem. Tento postup opakujeme, dokud nenalezneme kandidáta, kterého se nám nepodaří zfalzifikovat. Případně může procedura selhat tím, že množina přípustných parametrů je prázdná.

6 Sestavení řízení z demonstrací

Připomeňme, že naší původní úlohou je sestavit zpětnovazebné stabilizující řízení (2) respektující cenový funkcionál (3). Samotný demonstrátor v tomto ohledu není dostačující, neboť řízení jím poskytnuté je definované přes řešení NLP úlohy a navíc není zpětnovazebné.

Poznamenejme, že zde uvedená konstrukce stabilizujícího řízení z trajektorií není novým konceptem a tento postup nalezneme v [13] a [10]. Uvažme tedy diskrétní trajektorii poskytnutou demonstrátorem a stav systému x poblíž začátku této trajektorie. Zpětnovazebné řízení získáme pomocí *stabilizace trajektorie*. Její základní myšlenkou je zkonstruování řízení, jehož cílem je sledovat předpočítanou trajektorii (x_k^*, u_k^*) pomocí zavedení časově proměnného lineárně-kvadratického modelu (LQR). Konstrukce tohoto modelu se provede nahrazením původní dynamiky systému (1)

$$\dot{x} = F(x) + G(x)u$$

dynamikou

$$\Delta\dot{x} = A(x^*(t), u^*(t))\Delta x + B(x^*(t), u^*(t))\Delta u, \tag{6}$$

kde $(x^*(t), u^*(t))$ je sledovaná optimální trajektorie systému, $\Delta x \equiv x - x^*$ a $\Delta u \equiv u - u^*$ a

$$A(x^*(t), u^*(t)) \equiv \left.\frac{\partial(F(x) + G(x)u)}{\partial x}\right|_{x=x^*(t), u=u^*(t)},$$

$$B(x^*(t), u^*(t)) \equiv \left.\frac{\partial(F(x) + G(x)u)}{\partial u}\right|_{x=x^*(t), u=u^*(t)}.$$

Neboť je ale výstupem z demonstrátoru diskrétní trajektorie (x_k^*, u_k^*) přejdeme k diskrétnímu modelu

$$\Delta x_{k+1} = (I + hA_k)\Delta x_k + hB_k\Delta u_k \equiv A'_k\Delta x_k + B'_k\Delta u_k, \tag{7}$$

kde jsme zavedli značení $A_k \equiv A(x_k^*, u_k^*)$ a $B_k \equiv B(x_k^*, u_k^*)$.

Náš cíl je nyní zkonstruovat řízení, které bude sledovat předpočítanou trajektorii, tedy splňující $\Delta x_k, \Delta u_k \to 0$. Pro jeho zkonstruování zvolíme následující diskrétní LQR problém

$$\min \sum_{k=1}^{N} \Delta x_k^T P \Delta x_k + \Delta u_k^T Q \Delta u_k \tag{8}$$

za podmínek

$$\Delta x_{k+1} = A'_k\Delta x_k + B'_k\Delta u_k.$$

Řešení tohoto problému je lineární zpětnovazebné řízení tvaru $\Delta u_k = -K_k\Delta x_k$, kde K_k je dáno řešením příslušné diferenční Riccatiho rovnice [2]

$$S_k = Q + A_k^T(S_{k+1} - S_{k+1}B_k(R + B_k^T S_{k+1}B_k)^{-1}B_k^T S_{k+1})A_k \tag{9}$$

$$K_k = (R + B_k^T S_{k+1}B_k)^{-1}B_k^T S_{k+1}A_k. \tag{10}$$

Naše diskrétní řízení pro stav x v čase k, tedy bude

$$u_k = u_k^* - K_k(x - x_k^*), \tag{11}$$

a řízení pro spojitý čas $t \in \langle hk, h(k+1)\rangle$ lze zvolit jako lineární interpolace hodnot u_k a u_{k+1}.

Nyní ještě zmiňme jak pokračovat v generování řízení na konci trajektorie, neboť výstup z demonstrátoru je z principu konečný. Neboť lze očekávat, že v této době již budeme dostatečně blízko rovnovážnému stavu, můžeme sestavit lineární model poblíž rovnovážného stavu

$$\dot{x} = A(0,0)x + B(0,0)u,$$

a za funkcionál ceny zvolit původní funkcionál ceny (3). Tímto máme standardní LQR problém s nekonečným horizontem řešitelný pomocí algebraické Riccatiho rovnice pro spojitý čas [7]:

$$0 = A(0,0)^T S + SA(0,0) - SB(0,0)R^{-1}B^T(0,0)S + Q, \tag{12}$$

$$K = R^{-1}B^T(0,0)S \tag{13}$$

s řízením $u(t) = -Kx(t)$.

Zbývá ještě zvolit jednu trajektorii (x_k^*, u_k^*) při přítomnosti více trajektorií a též určit v jaké části trajektorie stabilizaci zahájit. Uvažme začátky dvou stejně dlouhých trajektorií (jednu z trajektorií můžeme doplnit nulovými stavy s nulovým řízením na potřebnou délku). Trajektorii můžeme zvolit pomocí matic K_1 získané z řešení Riccatiho rovnic (10), neboť hodnota $\Delta x_1^T K_1 \Delta x_1$ vyjadřuje cenu stabilizace pro diskrétní model (7) vzhledem k funkcionálu (8) [2]. Můžeme tedy vybrat tu trajektorii, která je v tomto smyslu levnější.

V tuto chvíli je ale třeba mít na paměti, že získané řízení bylo zkonstruováno na základní diskrétní aproximace spojité dynamiky systému a dále bylo využito lokální linearizace tohoto modelu. Je tedy snadno představitelné, že pro dostatečně vzdálený stav od předpočítaných trajektorií nebude toto řízení stabilizující.

7 Falzifikace stability získaného řízení

K vyšetřování řízení zkonstruované v části 6 využijeme CLF, kterou jsme získali na počátku našeho algoritmu. Zvolme tedy stav systému a napočítejme pro něj řízení získané stabilizací podle „nejbližší“ trajektorie a ověřme, že způsobuje pokles CLF. Tento postup však odpovídá falzifikaci řízení, při kterém budeme pro každý stav v každém čase napočítávat „nejbližší“ trajektorii a příslušné řízení. Toto však jistě není praktické.

Na druhém extrému leží nasimulování a kontrola celého řízení podél jedné na začátku vybrané trajektorie. To však víceméně odpovídá postupu v [10] a do značné míry se ztrácí benefit získání CLF, neboť ze simulace můžeme přímo vyhodnotit výsledný stav systému. Jako ideální kompromis se zdá být nasimulování pouhé částí trajektorie a kontrola této části vzhledem k napočítané Lyapunově funkci. To vlastně odpovídá falzifikaci řízení, při kterém bychom po určité dané době provedli přehodnocení cílované trajektorie.

Samotná falizifikace lze opět chápat jako hledání optima, v tomto případě maxima funkce

$$\max_{t\in\langle 0,t_{\max}\rangle} L(x(t),p)^T(F(x(t)) + G(x(t))u(x(t))),$$

avšak je třeba mít na paměti, že tato funkce má více lokálních maxim a ze způsobu napočítávání řízení obsahuje skokové nespojitosti. Z tohoto důvodu je prostor stavů v implementaci prozkoumáván globálním optimalizačním algoritmem nevyžadujícím napočítávání derivací funkce.

Nalezneme-li protipříklad, vygenerujeme z tohoto bodu novou trajektorii pomocí demonstrátoru, nalezneme novou kompatibilní CLF a získáme nové řízení. Takto budeme doplňovat trajektorie, dokud nedojde k selhání ve smyslu nenalezení vhodné CLF, či dokud nebudeme schopni nalézt protipříklad.

8 Celkový přehled algoritmu

1. Počáteční systém trajektorií $\mathcal{T}$ obsahující demonstrace $(\mathcal{X},\mathcal{U})$ získaných demonstrátorem a systém lineárně parametrizovaných kandidátů na CLF $\{L(x,p)\}$.

2. Do nalezení stabilizujícího řízení nebo selhání procedury:

   (a) Nauč se CLF L na demonstracích z $(\mathcal{X},\mathcal{U})$ hledáním kandidáta $L(x,p)$ splňující podmínky:

      i. $L(x,p) > 0$ pro všechny demonstrace (x,u) nenulové.

      ii. $\nabla L(x)^T(F(x)+G(x)u) < 0$ pro všechny demonstrace (x,u) nenulové.

   a následně hledej protipříklady řešením optimalizačních úloh

      i. $\min_x L(x,p)$,

      ii. $\max_x \nabla L(x,p)^T(F(x)+G(x)u_S(x))$, kde u_S je řízení dané Sontagovou formulí (4).

   Doplňuj do $\mathcal{T}$ a $(\mathcal{X},\mathcal{U})$ případné nalezené protipříklady s využitím demonstrátoru.

   (b) Na aktuálním $\mathcal{T}$ sestroj řízení u pomocí linearizace dle sekce 6.

   (c) Hledej protipříklady k tvrzení, že je získané u stabilizující řešením optimalizační úlohy

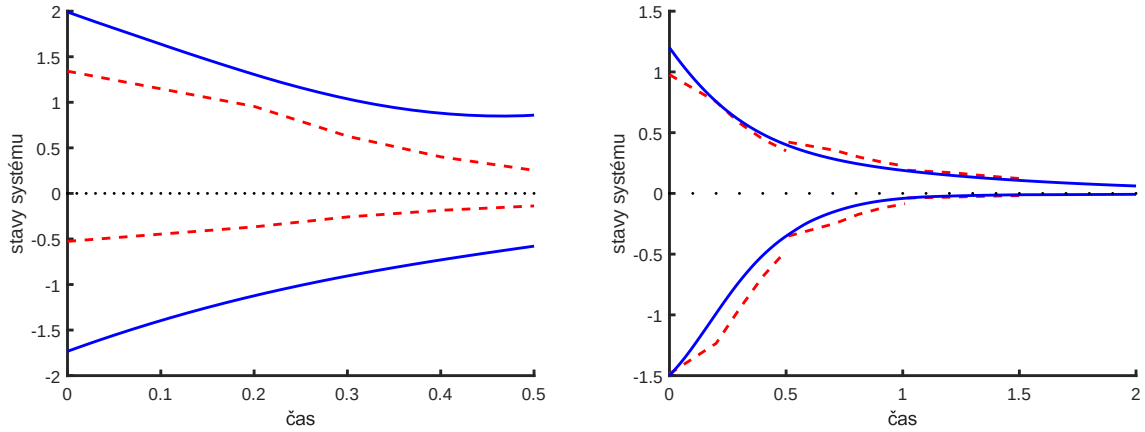
$$\max_{t\in\langle 0,t_{\max}\rangle} L(x(t),p)^T(F(x(t))+G(x(t))u(x(t))),$$

   kde $(x(t)),u(x(t)))$ jsou nasimulované trajektorie po zvolenou dobu $t_{\max}$, viz sekce 7. Pokud nalezen protipříklad x, získej z demonstrátoru trajektorii z x a doplň do $(\mathcal{X},\mathcal{U})$ příslušné demonstrace. Následně opakuj celý postup (a)–(c) znovu.

9 Příklad

Uvažme následující systém afinní v řízení

$$\begin{aligned}\dot{x_1} &= -x_1 + x_2\\ \dot{x_2} &= \frac{1}{2}x_1 - \frac{1}{2}x_2(1-\cos 2x_1+2)^2 + (\cos 2x_1+2)u\end{aligned}$$



(a) (b)

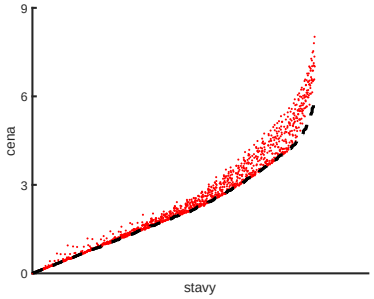
Obrázek 1: Protipříklad při kontrole řízení (a) a příklad finální získané trajektorie (b): plné čáry označují nasimulovanou trajektorii obou stavů, přerušované cílovanou trajektorii, tečkovaná čára označuje rovnovážný stav

a nechť jsou matice P a Q v (3) zvoleny jako jednotkové. Pro aplikaci představeného algoritmu je nejprve nutné zvolit vhodný demonstrátor. Pro tento účel užijeme balíček CasADi [1] implementovaný pro MATLAB specializovaný pro řešení úloh nelineárního programovaní plynoucích z problému optimálního řízení. Za diskretizaci dynamiky systému uvážíme Hermiteovu-Simpsonovu kolokační metodu [6]. Jako aproximaci nekonečného horizontu zvolíme délku časového horizontu jako minimálně 4 s tím, že trajektorii z demonstrátoru akceptujeme, pokud je norma posledního stavu menší než 0.1 (jinak je časový horizont prodloužen). Délka časového kroku byla zvolena jako 0.1.

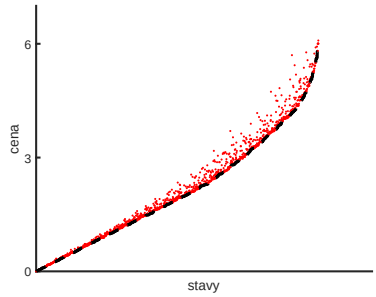
Jako parametrizovaný systém kandidátu na CLF zvolíme ryze kvadratické polynomy a omezíme se na množinu stavů $\langle -2, 2\rangle \times \langle -2, 2\rangle$ (pro tento systém nedochází při stabilizaci k opuštění této množiny). Pro nalezení kandidáta využijeme nástroj YALMIP [8] s vnitřním řešičem SDPT3 [14] pomocí kterého nalezneme střed v tomto případě vepsané elipsy o maximálním objemu pomocí řešení příslušné úlohy semidefinitního programování. Časový krok pro přepínání zvolíme jako 0.5, který tedy i hraje roli délky simulací systému pro ověřování. Samotné simulace provedeme pomocí metody RKF45. Pro samotnou falzifikaci využijeme v případě CLF matlabovskou funkci fmincon, pro falzifikaci trajektorie diferenciální evoluci.

Ilustraci postupu můžeme vidět na obrázku 1a, kde je uveden příklad zamítnuté vygenerované trajektorie kvůli nekompatibilitě s nalezenou CLF. Z obrázku vidíme, že skutečně jeden stav přestává být v závěrečné fázi trajektorie stabilizován. Na obrázku 1b vidíme příklad kompletní trajektorie systému včetně přepínáni cílované trajektorie pro jeden vybraný stav.

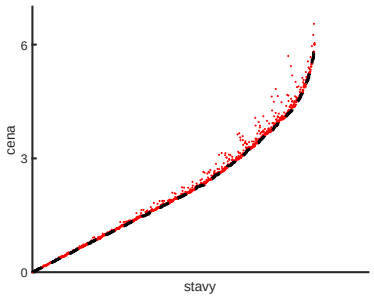
Falzifikace řízení byla ukončena po přidání 5 trajektorií. Po té se již nepodařilo najít protipříklad nekompatibilní s nalezenou CLF. Můžeme si povšimnout, že přidání dalších trajektorií (na základě maximalizace vzdálenosti od již přítomných trajektorií) přibližuje řízení k optimálnímu řízení pro tento příklad, které je určeno cenou $V^*(x) = \frac{1}{2}x_1^2 + x_2^2$. Toto vidíme na obrázku 2, kde bylo cena řízení nasimulována přibližně z 1500 stavů.



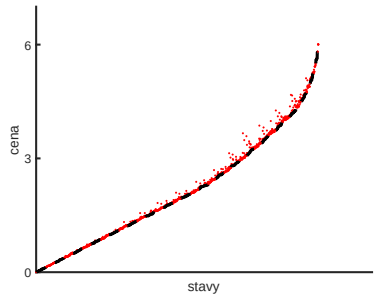
(a) 5 trajektorií



(b) 10 trajektorií



(c) 15 trajektorií



(d) 20 trajektorií

Obrázek 2: Cena řízení dle počtu trajektorií, body označují cenu získaného řízení pro vybrané stavy (seřazené dle optimální ceny), přerušovaná čára vyznačuje optimální cenu

10 Závěr

V tomto textu byl představen postup na generování řízení pomocí demonstrací vygenerovaných demonstrátorem, který je založen na přímém řešení úlohy optimálního řízení. Toto řízení je ověřováno ve smyslu stabilizace falzifikací pomocí konstrukce Lyapunových funkcí pro systém s řízením. Postup byl demonstrován na jednoduchém v řízení afinním problému. Pro další analýzu je nutné testování na dalších složitějších problémech. Již s předstihem je předvídatelným nedostatkem značná závislost konstrukce řízení na demonstrátoru. Na druhou stranu řízení zkonstruované demonstrátorem respektuje cenový funkcionál, na rozdíl od obvyklého postupu konstrukce stabilizujícího řízení pomocí Sontagovy formule. V budoucnu je zejména záhodno modifikovat postup tak, aby dokázal efektivně fungovat i pro demonstrátor, který sám o sobě neposkytuje vždy stabilizující řízení.

Literatura

[1] J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl. *CasADi – A software framework for nonlinear optimization and optimal control*. Mathematical Programming Computation **11** (2018).

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, (2005).

[3] J. T. Betts. *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*. SIAM, (2010).

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, (2004).

[5] R. A. Freeman and J. A. Primbs. *Control Lyapunov functions: new ideas from an old source*. Proceedings of 35th IEEE Conference on Decision and Control **4** (1996), 3926–3931.

[6] C. R. Hargraves and S. W. Paris. *Direct trajectory optimization using nonlinear programming and collocation*. Journal of Guidance, Control, and Dynamics **10** (1987), 338–342.

[7] D. Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, (2011).

[8] J. Löfberg. Yalmip : A toolbox for modeling and optimization in matlab. In 'In Proceedings of the CACSD Conference', Taipei, Taiwan, (2004).

[9] H. Ravanbakhsh and S. Sankaranarayanan. *Learning control Lyapunov functions from counterexamples and demonstrations*. Autonomous Robots **43** (2019), 275–307.

[10] P. Reist, P. Preiswerk, and R. Tedrake. *Feedback-motion-planning with simulation-based LQR-trees*. The International Journal of Robotics Research **35** (2016), 1393–1416.

[11] E. D. Sontag. *A 'universal' construction of Artstein's theorem on nonlinear stabilization*. Systems & Control Letters **13** (1989), 117–123.

[12] E. D. Sontag. *Mathematical Control Theory*. Springer-Verlag New York, (1998).

[13] R. Tedrake, I. R. Manchester, M. Tobenkin, and J. W. Roberts. *LQR-trees: Feedback motion planning via sums-of-squares verification*. The International Journal of Robotics Research **29** (2010), 1038–1052.

[14] K. Toh, M. Todd, and R. Tutuncu. *Sdpt3 — a Matlab software package for semidefinite programming*. Optimization Methods and Software **11** (1999), 545–581.

Density-Approximating Neural Network Models for Anomaly Detection Using Non-Uniform Auxiliary Training Sets

Martin Flusser

4th 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

Tomáš Pevný, Department of Computer Science
Faculty of Electrical 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 non-anomalous data . This makes the task difficult because the character of the anomalous data is unknown when the model is trained.

We propose an adaptive approach for density-approximating neural network models, the alternative use of neural models in anomaly detection. Traditionally, in anomaly detection context the common use of neural models is in form of auto-encoders. Through the use of auto-encoders the true anomality is proxied by reconstruction error. Auto-encoders often perform well but do not guarantee to perform as expected in all cases. A popular more direct way of modeling anomality distribution is through k-Nearest Neighbor models. Although kNN can perform better than auto-encoders in some cases, their applicability can be seriously impaired by their space and time complexity especially with high-dimensional large-scale data. The density-approximating neural networks [10] model the distribution imposed by kNN using. It was proved that such neural models are capable of achieving comparable accuracy to kNN while reducing computational complexity by orders of magnitude on benchmark data form UCI repository. However, the original method could be insufficient in some cases as shown in this paper.

We propose an adaptive approach that provides better accuracy specifically for more complex data. We evaluate the proposed idea against the non-adaptive approach and standard kNN on data from computer network traffic provided by Cisco Systems company and show that in majority of cases it is possible to improve on accuracy or computational cost.

Keywords: Anomaly detection, 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 této práci je prezentován adaptivní přístup pro neurální aproximační metodu, která je alternativou k tradičně používaným metodám pomocí neuronových sítí v podobě autoencoderů. Nevýhoda autoencoderů je ta, že anomalitu nemodelují přímo, ale jen pomocí rekonstrukční chyby. Přímou cestou jak modelovat anomalitu je například použití kNN. O kNN je známo že funguje pro detekci anomálií velmi dobře a zatím nebyla nalezena metoda, která by obecně fungovala lépe. Velkou nevýhodou kNN je však jeho výpočetní náročnost, která brání použití pro mnoho reálných aplikací. V minulosti se ukázalo, že neurální model pro aproximaci hustoty, je schopen se kvalitativně vyrovnat kNN a zároveň výrazně zredukovat výpočetní náročnost. Tato metoda byla v minulosti evaluována na veřejných datasetech z UCI repositáře avšak na datasetu z oboru komunikace na počítačových sítích nedosahuje v původně publikované formě požadovaných výsledků, jak prezentujeme v tomto článku. Adaptivní přístup, který je představen v tomto článku výrazně vylepšuje přesnost dříve publikované metody a to zejména pro komplikované datasety. Toto vylepšení je evaluováno na datech z provozu na počítačových sítích poskytnutých společností Cisco Systems a je ukázáno, že adaptivní přistup má lepší výsledky než původní metoda a také že na rozdíl od původní metody dorovnává výsledky kNN.

Klíčová slova: Detekce anomálií, 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.

Anomaly detection (AD) is gaining on importance with the massive increase of data we can observe in every domain of human activity. In many applications the goal is to recognize objects or events of classes with unclear definition and missing prior ground truth, while the only assumed certainty is that these entities should be different from what we know well. The problem can thus be seen as the problem of modeling what is common, and then identifying outliers.

Applications of anomaly detection are extensive. Anomaly detection is inherent in cyber security, is successfully applied in industrial quality control, banking and credit card fraud detection, in medicine it can help raise alarms when a patient's condition deteriorates, etc.

Anomaly detection as a general problem has been widely studied as we overview in Section 1.1. Current state of the art is, however, less satisfactory than in supervised learning. Specifically, the recent rapid advances in neural networks (for overview see, e.g., [16], [7], [13]) seem to not have been replicated as successfully in anomaly detection.

The primary neural model use in anomaly detection is through *auto-encoders* (AE). Auto-encoders, however, do not model distribution of anomalies, they optimize a proxy criterion, usually in form of reconstruction error. This fact can limit the success of AEs in some problem areas.

Among traditional anomaly detection principles the *k-nearest neighbor* (kNN) remains among the best performing models. Distance-based detectors directly model the density

but can become computationally expensive or even prohibitive in on-line and embedded systems.

Density-approximating neural networks models take use of distance-based kNN principle to enable training of neural models with multiple potential advantages: better robustness against noise as well as low computational complexity leading to high detection speed - an important parameter especially in on-line and embedded anomaly detection applications.

In this paper we propose an adaptive approach for density-approximating neural networks models and evaluate its performance for complex real data of computer network traffic.

The paper is structured as follows: in Section 1.1 we review existing anomaly detection methodology, in Section 2 we introduce the proposed method, in Section 3 we cover the experimental evaluation of the proposed method and comparison to kNN in Sections 4 and 5 we provide discussion and conclusion.

1.1 Anomaly detection

Anomaly detection is also known as one-class classification. The goal is to detect a sample that is somehow different from expected patterns or other observations, without knowing the exact definition of *different*. Hence, anomaly detection techniques focus on modeling what is expected, and subsequently to mark as anomaly anything sufficiently different from the expected.

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.

There is a number of methods for anomaly detection the survey of which is given, e.g., in [5], [18], [21]. This paper focuses on nearest neighbor based techniques [15] and neural models and consequently investigates the question of how to find synergy between both. Nearest neighbor techniques are beneficial for their performance (under certain conditions) and adaptability to various data types. Their computational complexity, however, grows rapidly with both the dimensionality and size of the training data. Supporting structures thus have been proposed especially in form of *k-d trees* [1, 11, 2] and *ball trees* [25, 2] to mitigate the problem. But the problem of complexity has thus not disappeared.

The standard anomaly detection knowledge base also includes *kernel PCA* methods [19], *kernel density estimation (KDE)* including *robust KDE* [14] and *one-class support vector machines (SVM)* [24] that all have been compared to and partly outperformed by neural models, see, e.g., [26]. Neural network models are used for anomaly detection in two different ways: 1) fully unsupervised, i.e., neural network is trained on the regular data only and produces anomaly score or any other similar metric which can be thresholded , 2) supervised to some extent, i.e., knowledge about possible outliers or other indirect information about anomality apart from the mere density is utilized during training (see, e.g., [4], [22], [23], [12], [27], [20]). In the following we consider only the standard approach to anomaly detector training where no additional information is assumed available apart from the unlabeled data.

2 Proposed Method

The proposed method aims to make nearest neighbor based anomaly detection efficient utilizing a neural network. The main idea is simple: train a neural network that estimates kNN score. The algorithm does it in two logical steps. First, it creates an auxiliary dataset covering the input space, and for each point of this auxiliary set a kNN anomaly score is computed. Then, this auxiliary dataset is used as a training set to train the neural network-based estimator.

Having the training set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^d$, $\forall i \in \{1, .., n\}$, let us denote $\mathbf{A}$ the auxiliary data set of m samples where $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m\}$, $\mathbf{a}_i \in \mathbb{R}^d$, $\forall i \in \{1, .., m\}$ and Y the vector of respective anomaly scores, where $Y = \{y_1, y_2, ..., y_m\}$, $y_i \in \mathbb{R}$), $\forall i \in \{1, .., m\}$. We will consider the size of the proposed neural network's hidden layers to be $d \cdot p$ where p is a parameter.

2.1 Auxiliary set construction

In this stage the auxiliary set $\mathbf{A}$ is computed from the training set $\mathbf{X}$. In addition to the *uniform auxiliary set computing* we introduce *adaptive auxiliary set computing* which covers the space in more efficient way and prevents learning errors stemming from the hard space boundary thresholding in the uniform set case.

2.1.1 Uniform auxiliary set construction

The idea from [10] is naïve as it attempts to cover the anomaly space uniformly on a rectangular subspace defined as smallest enclosing hyper-block that contains all points in the input data space.

1. A bounding hyper-block of $\mathbf{X}$ is observed. Such hyper-block is defined with the vector of lower bounds $\mathbf{h}_l$ and upper bounds $\mathbf{h}_u$ such that $\mathbf{h}_l^{(j)} \leqslant \mathbf{x}_i^{(j)} \leqslant \mathbf{h}_u^{(j)} \quad \forall i \in \{1, .., n\} \quad \forall j \in \{1, .., d\}$ where $\mathbf{x}_i^{(j)}$ represents j-th element of i-th vector from $\mathbf{X}$

2. The hyper-block is filled with randomly generated and uniformly distributed samples $\{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m\}$. By default we consider uniform random sampling. Note that the choice of m for concrete problem may depend on n and d (see also Sec. 3.2.2).

3. The anomaly score vector Y is constructed so that for each auxiliary sample $\mathbf{a}_i, i \in \{1, .., m\}$ the respective $y_i \in Y$ is computed as k-Nearest Neighbor mean distance $G(\cdot)$:

$$y_i = G(\mathbf{a}_i) = \frac{1}{k} \sum_{j=1}^{k} D_j(\mathbf{a}_i) \tag{1}$$

   where $D_j(\mathbf{a}_i)$ represents the j-th smallest distance of $\mathbf{a}_i$ to samples from $\mathbf{X}$. Note that the number of neighbors k is a parameter [28, 17].

2.1.2 Adaptive auxiliary set construction

Uniform auxiliary as defined above set is sub-optimal due to multiple reasons. Clearly the distribution of points in uniform auxiliary set does not reflect varying importance of various regions in the auxiliary space; uniform auxiliary set can easily waste sampled points in regions of no importance while lacking coverage in dense complicated modes. Another problem is the definition of the bounding hyper-block; its hard boundaries can lead to misrepresentation of true distribution of anomalies. It may and does happen that due to sampling the input data represent distributions that in fact should be modeled way outside the boundaries of the hyper-block that rely too tightly on the particular sampling represented in the input data.

We propose to construct auxiliary data set adaptively to reflect the distribution in input data. This is achieved by generating auxiliary samples according to a modified Parzen estimate of the input density. No bounding hyper-block is thus needed, while the auxiliary samples now become more frequent in areas of more detail. The resulting auxiliary data set is thus expectably significantly better than the uniform one of the same size (nr. of samples).

1. Optimal variance h for Parzen window approximation of $\mathbf{X}$ is discovered.

2. The auxiliary set $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m\}$ is created as realization of Parzen approximated distribution of $\mathbf{X}$ as follows: We iterate over samples of $\mathbf{X}$ and create $\mathbf{a}_i = \mathbf{x}_i + \mathcal{N}(0,\, h \cdot k_{var})$ where k_{var} (variance multiplicative coefficient) is a parameter. Typically $m > n$ thus one sample form $\mathbf{X}$ generates more samples in $\mathbf{A}$:

   $$\forall i \in \{1, ..., m\} :\ \mathbf{a}_i = \mathbf{x}_{i \mod n} + \mathcal{N}(0,\, h \cdot k_{var})$$

   Note that the choice of m and k_{var} for concrete problem may depend on n and d (see also Sec. 3.2.2).

3. The anomaly score vector Y is constructed in the same way as for uniform auxiliary set. (see Sec. 2.1.1, step 3)

2.2 Training of the model

The feed forward multi-layer neural network (see Fig.1) is trained with $\mathbf{A}$ and Y to be able to predict the anomaly score. In other words, the input vector $\mathbf{a}_i \in \mathbb{R}^d$ is projected to $y'_i \in \mathbb{R}$ as follows:

$$y'_i = f_\theta(\mathbf{a}_i) = f^{(4)}_{\theta^{(4)}}(f^{(3)}_{\theta^{(3)}}(f^{(2)}_{\theta^{(2)}}(f^{(1)}_{\theta^{(1)}}(\mathbf{a}_i)))) \tag{2}$$

where $f^{(j)}_{\theta^{(j)}}$ represents the j-th layer of the NN and the layer propagation is defined as:

$$f^{(j)}_{\theta^{(j)}}(\mathbf{a}_i) = c(\mathbf{W}^{(j)}\mathbf{a_i} + \mathbf{b}^{(j)}) \tag{3}$$

thus $f^{(j)}$ is parameterized by $\theta^{(j)} = \{\mathbf{W}^{(j)}, \mathbf{b}^{(j)}\}$, c is an activation function, $\mathbf{W^{(j)}}$ is a weight matrix and $\mathbf{b^{(j)}}$ is a bias vector of the j-th layer.

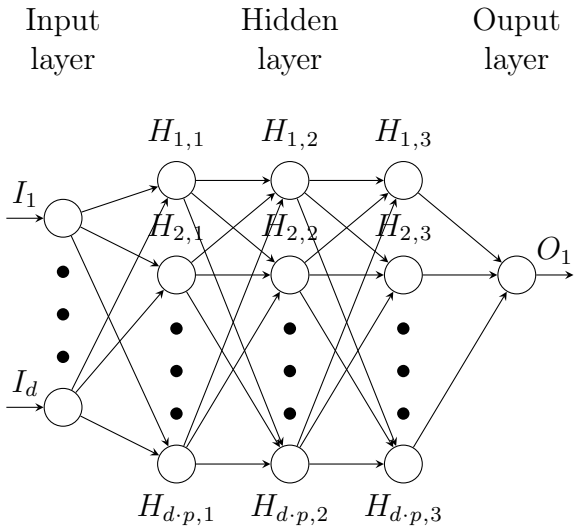


Figure 1: Structure of the utilized network

The parameters of the model are optimized with $\mathbf{A}$ and Y such that the average loss function is minimized:

$$\theta^* = \arg\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} L\Big(y_i, y_i'\Big) \tag{4}$$

where L represents a loss function.

3 Experimental Evaluation

The main aim of the experiment is to compare the two mentioned approaches to standard kNN based anomaly detection and to discover possible advantage of adaptive approach in comparison to uniform approach. The experiment is performed on real data from computer network traffic. As a metric we use area under the curve of receiver operating characteristics (AUC of ROC) [3]. [1]

3.1 Data Set

Our aim is to operate on real data form computer network traffic. The dataset was provided by Cisco Systems company and it describes connections between users and servers. Originally, the datset is multiclass thus to create an AD dataset we have adopted the experimental protocol of Emmott [9], who has introduced a methodology of creating general AD benchmarking sets using multi-class data.

One selected class is used to form the non-anomalous data and some of the others are used as source of anomalies. The concrete choice of anomaly representing classes leads up to four different datasets of four levels of detection' difficulty. Emmott demonstrated the

[1] The utilized hardware for this experiment is provided by the supercomputer services of the computing and information center at the Czech Technical University and Cognitive Research at Cisco Systems, Prague.

utility of such approach by creating a number of carefully selected sets and using them to evaluate performance of six popular AD method. There is a large number of various multi-class data sets usually based on a real-world data in the UCI repository hence the constructed benchmark sets provide a reasonable reality check. In 2014 Dau considered Emmott's methodology as the most advanced [6].

To give more insight into method's performance under various conditions the anomalous data are grouped according to their difficulty. We thus perform our evaluation on *easy* , *medium* , *hard* and *very hard* problems. The set has 222 455 samples and dimension 10.

3.2 Evaluation Setup

To construct training and testing sets, in all cases random sampling is used such that 75 % of normal (non-anomalous) samples is used for training and the rest 25 % for testing. The anomalous samples are only used for testing. The score is measured with AUC of ROC [3] as is common in literature. The advantage of this metric is the independence on specific thresholding.

3.2.1 k-Nearest Neighbors Setup

To evaluate kNN accuracy we compute AUC according to the anomaly score obtained as mean distance $G(\cdot)$ introduced in Equation (1).

The optimal choice of the parameter k which is essential for kNN is not addressed in this paper. However, we observed $k = 5$ as the best performing thus it is used for all presented experiments. Remark: note that the proposed method is applicable for any k.

3.2.2 Proposed Method Setup

To evaluate the accuracy of the proposed model we compute AUCs using the neural network introduced in Section 2.2. The method is subject to parametrization: its performance can be affected by the properties of auxiliary data set as well as by the standard neural model parametrization (number of layers, number of neurons in layers etc).

We fixed the auxiliary set construction parameters for all experiments as follows. We fixed $k = 5$ in kNN used for auxiliary data set construction to get results comparable to the standalone kNN anomaly detector. The auxiliary data set is constructed as described in Section 2.1.1 with the total number of auxiliary samples set to $m = n \cdot d \cdot l$, where l is set to 18. The choice of the parameter l is empirical and reflects a trade-off between model accuracy and computational complexity of the training.

ReLU ($f(x) = \max(0, x)$) activation function is used for all neurons (except input). Size of batch is set always to 80. We opted for a simple meta-optimization of neural model parameters so as to avoid the worst local optima. The same procedure is applied across all benchmark data. For this purpose we train for each training data set multiple models, to eventually retain the version with best loss function result. The variation across training runs consist in: 2 or 3 hidden layers, hidden layer size $3d$ or $5d$, multiple random weight initialization, number of iterations thresholded by six values between 15000 and 700000.

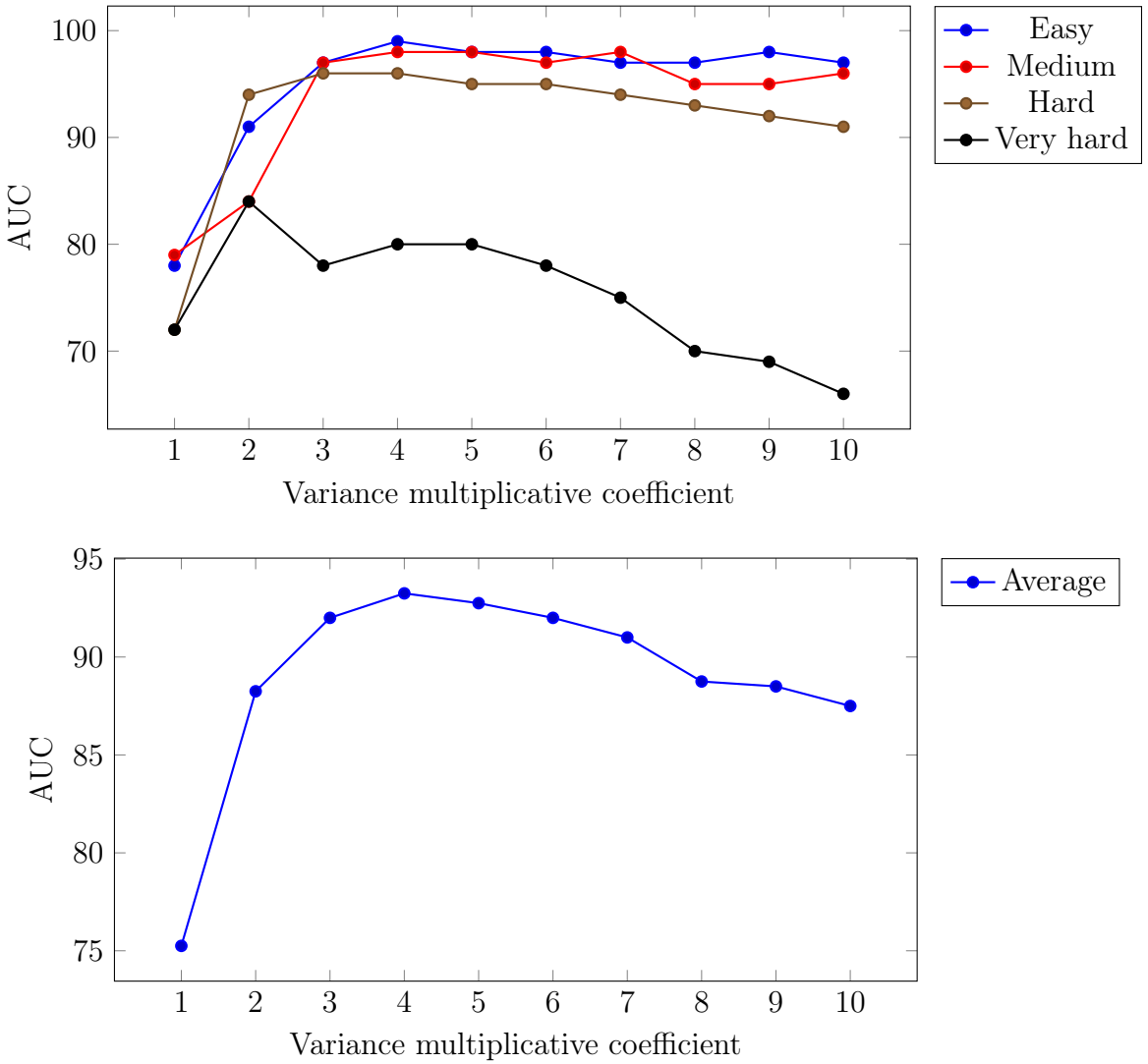


Figure 2: Top: Performance's dependence on parameter k_{var} for each difficulty. The chart indicates that optimal value could vary on the problem difficulty. The *very hard* problem reach maximal score for $k_{var} = 2$ for which even *hard* performs better than *easy* and *medium*. Except *very hard*, the scores reach the maximum near $k_{var} = 4$. With growing k_{var} above 4 the scores decrease rapidly for harder problems. Bottom: Performance's dependence on parameter k_{var} averaged over all difficulties. It is apparent that the optimal choice is $k_{var} = 4$

3.2.3 Hyper-parameter tuning for adaptive auxiliary set computing

The adaptive approach utilize parameter k_{var} (variance multiplicative coefficient) which is essential to correct the variance obtained by the Parzen window coverage that surprisingly must be scaled to become efficient. Full analysis of various coefficients for each difficulty separately is given in Fig. 2. However, the optimal parameter was selected over all difficulties as maximum of the average score.

3.3 Detection Accuracy Results

Table 1: *proposed method* versus *kNN*, grouped by problem difficulty.

	Easy	Medium	Hard	V. Hard	Win	Avg
*k*NN	96,0	94,9	**96,2**	**90,8**	**2**	**94,5**
Uniform	94,0	90,3	79,4	65,9	0	82,4
Adaptive	**98,5**	**97,7**	95,9	80,0	**2**	93,0

Assessing results of three methods over multiple datasets (diffculties) can be done in multiple ways [8]. We provide the achived AUC accuracies in table 1, each column covering one problem difficulty level with highlighted best score for each problem. The proposed method with uniform approach achieved lowest score for each problem. However the adaptive approach outperforms *k*NN at *easy* and *medium* level. Two overall comparison methodologies such as counts of wins and averages over the datasets are provided in the table.

The results show that the adaptive approach outperforms the uniform approach at most difficulty levels. The results do not show significant difference between the proposed method with adaptive approach and coventional *k*NN with exception of the very hard anomalies.

To summarize, the introduction of adaptive auxiliary set has enabled considerable improvement of accuracy achieved by the neural model on a real world dataset that can be considered challenging in general. The computational complexity in the application phase is not affected by switching to the adaptive approach; for details we refer to the complexity analysis given in [10].

4 Discussion

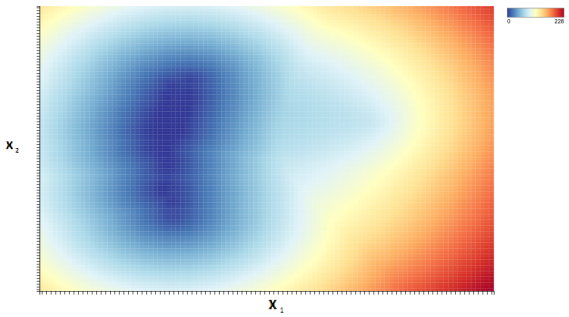


Figure 3: Anomaly scores on a 2D projection of the utilized data set inferred by distance to k-nearest neighbours. Warmer color depicts higher anomaly.

To give more insight into how the proposed model replicates *k*NN-induced anomaly distribution we provide heat-maps on 2D projection of the utilized dataset. Fig.3 depicts anomaly score distribution obtained by *k*NN. Fig.4 demonstrates the anomaly score

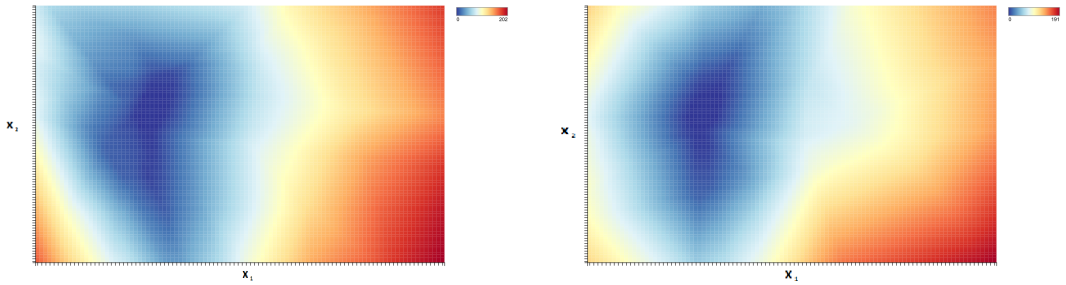


Figure 4: Anomaly scores on a 2D projection of the utilized data set. Anomalousness inferred by the proposed model. Non-adaptive approach is on the left side, adaptive on the right side. Warmer color depicts higher anomaly.

distribution imposed by the proposed model for both the non-adaptive and adaptive approach.

To construct the heat-maps the dataset were transformed into 2D space using PCA. The respective anomaly in each pixel position is marked by color on a scale from blue (lowest anomaly) to red (highest anomaly).

Note also that we have not exhausted all parametrization options of the neural network when comparing the proposed model to kNN. It should be also noted that the main idea behind our proposed method does not actually depend on neural networks. Once an auxiliary set is constructed, it should be possible to apply any predictor capable of learning from samples with labels from $\langle 0, 1 \rangle$.

In this experiment, we assume there is the only optimal value for parameter k_{var} even though separate tuning for each problem would reach to better score of the adaptive approach. Our experiment is based on real simulation where no information about the character of tested data is available. However there is an opportunity for further research about how to tune the parameter with respect to what types of anomalous data we focus on.

The accuracy of the proposed method thus depends crucially on the number and distribution of auxiliary samples. In the present paper we assume only the constant number of axiliary samples for both approaches of the proposed experiment (cf. Section 3.2.2). Expectably the accuracy of the proposed model can be improved further by optimization of the auxiliary set size with respect to data set properties. This is the subject of our next effort.

5 Conclusion

We propose an novel adaptive approach for density-approximating neural network models for anomaly detection. We take use of distance-based k-Nearest Neighbor principle to enable unsupervised training of neural networks that directly model the density of anomaly values. This is in contrast to most neural networks where anomality is modeled indirectly through reconstruction error or other proxy criterion. The non-adaptive approach has been shown to perform well [10] while using only a naïve auxiliary set construction; in this paper we use the adaptive approach that better corresponds with the nature of the

input data and that is able to achieve significantly better coverage with the auxiliary set. This is crucial for successful modeling of non-trivial data sets, such as real data from network traffic, where the uniform sampling appears insufficient.

We compare the proposed approach's accuracy to kNN on an set of real data. To obtain robust results we defined meta-optimization of parameters for both approaches of the compared neural models. The evaluation shows that the proposed approach exhibits multiple advantages. When compared to the uniform density approach it often provides better accuracy. When compared to kNN it provides comparable accuracy with principally lower computational complexity - an important property especially in on-line and embedded anomaly detection applications.

References

[1] J. L. Bentley. *Multidimensional binary search trees used for associative searching.* Communications of the ACM **18** (1975), 509–517.

[2] A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In 'Proceedings of the 23rd international conference on Machine learning', 97–104. ACM, (2006).

[3] C. D. Brown and H. T. Davis. *Receiver operating characteristics curves and related decision measures: A tutorial.* Chemometrics and Intelligent Laboratory Systems **80** (2006), 24–38.

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

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

[6] H. A. Dau, V. Ciesielski, and A. Song. Anomaly detection using replicator neural networks trained on examples of one class. In 'Asia-Pacific Conference on Simulated Evolution and Learning', 311–322. Springer, (2014).

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

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

[9] 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.

[10] M. Flusser, T. Pevný, and P. Somol. *Density-approximating neural network models for anomaly detection.* ACM SIGKDD workshop on outlier detection de-constructed (8 2018). London, United Kingdom. `https://www.andrew.cmu.edu/user/lakoglu/odd/accepted_papers/ODD_v50_paper_19.pdf` or: `goo.gl/73yvmG`.

[11] J. H. Friedman, J. L. Bentley, and R. A. Finkel. *An algorithm for finding best matches in logarithmic expected time.* ACM Transactions on Mathematical Software (TOMS) **3** (1977), 209–226.

[12] 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).

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

[14] J. Kim and C. D. Scott. *Robust kernel density estimation.* Journal of Machine Learning Research **13** (2012), 2529–2565.

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

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

[17] C. R. Loader. *Local likelihood density estimation.* Ann. Statist. **24** (08 1996), 1602–1618.

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

[19] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch. Kernel PCA and de-noising in feature spaces. In 'Advances in neural information processing systems', 536–542, (1999).

[20] 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).

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

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

[23] 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 **35** (2005), 302–312.

[24] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. *Estimating the support of a high-dimensional distribution.* Neural computation **13** (2001), 1443–1471.

[25] J. K. Uhlmann. *Satisfying general proximity/similarity queries with metric trees.* Information processing letters **40** (1991), 175–179.

[26] 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).

[27] 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).

[28] M. Zhao and V. Saligrama. Anomaly detection with score functions based on nearest neighbor graphs. In 'Advances in neural information processing systems', 2250–2258, (2009).

Matematické modelování transportu komponent vícesložkové směsi v porézním prostředí*

Petr Gális

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

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

Abstract. This work deals with higher-order numerical scheme for multicomponent flow of mixture in porous media. Mathematical model uses Darcy law to describe velocity field, volume balance equations for the components of the mixture and volume balance equation with prescribed initial and boundary conditions. Problem is solved via iterative IMPEC (implicit pressure, explicit concentration) method with mixed-hybrid finite element method to discretize Darcy's velocity using higher-order Raviart-Thomas element and discontinuous Galerkin method to discretize transport and pressure equations. In the numerical experiment of the one-dimensional flow the functionality of the scheme is demonstrated including its presumed order of convergence.

Keywords: Compositional flow, Mixed-hybrid finite element method, Discontinuous Galerkin method

Abstrakt. V této práci se zabýváme numerickým schématem vyššího řádu přesnosti pro proudění vícesložkové směsi v porézním prostředí. Matematický model využívá popisu rychlosti pomocí Darcyho zákona, bilančních rovnic pro komponenty směsi a rovnice pro objemovou bilanci s počátečními a okrajovými podmínkami. Problém řešíme iterativní metodou IMPEC (implicit pressure, explicit concentrations) v kombinaci se smíšenou hybridní verzí metody konečných prvků za použití vyššího Raviartova-Thomasova prvku pro diskretizaci Darcyho zákona a nespojitou Galerkinovou metodou pro diskretizaci transportních rovnic a rovnice pro tlak. V numerickém experimentu jednodimenzionálního proudění je demonstrována funkčnost schématu včetně jeho předpokládaného řádu přesnosti.

Klíčová slova: Kompoziční proudění, Smíšená hybridní metoda konečných prvků, Nespojitá Galerkinova metoda

1 Úvod

Tradiční přístup k modelování transportu směsí je použití metod nízkých řádů pro aproximaci příslušných rovnic. To může vést k nutnosti řešit problém na velmi jemných sítích, což zejména při použití neimplicitních řešičů vede k jejich nestabilitě při větších časových

*Tato práce byla podpořena grantem SGS17/194/OHK4/3T/14.

krocích. V této práci implementujeme novou iterační metodu [3] spolu s metodami vyšších řádů, přičemž uvidíme, že výsledné numerické řešení je již při použití hrubé sítě relativně přesné jak kvalitativně, tak kvantitativně. V následujícím textu představíme základy modelování proudění vícesložkové směsi v porézním prostředí a podrobněji rozebereme odvození příslušných rovnic pro potřeby metod konečných prvků a diskrétní Galerkinovy metody, přičemž se omezíme na výklad v jedné dimenzi.

2 Matematický model

2.1 Transportní rovnice

Předpokládejme stlačitelné 1D proudění jednofázové směsi o n_c komponentách v porézním prostředí s danou porozitou ϕ při konstantní teplotě T. Zanedbáme-li difuzi, transport komponent směsi lze popsat následujícími bilančními rovnicemi hmoty

$$\frac{\partial\left(\phi(\mathbf{x})c_i(\mathbf{x},t)\right)}{\partial t} + \nabla\cdot\left(c_i(\mathbf{x},t)\mathbf{v}(\mathbf{x},t)\right) = f_i(\mathbf{x},t), \quad i = 1,...,n_c, \tag{1}$$

kde c_i je celková molární koncentrace i-té komponenty, f_i je zřídlový člen i-té komponenty a $\mathbf{v}$ je rychlostní pole. Rychlostní pole zde popisujeme pomocí Darcyho zákona s absencí gravitačního pole

$$\nabla p = -\mu\mathbf{k}^{-1}\cdot\mathbf{v}, \tag{2}$$

kde $\mu[kg\cdot m^{-1}\cdot s^{-1}]$ je viskozita a $\mathbf{k} = \mathbf{k}(\mathbf{x})$ je permeabilita prostředí. Z bilance objemu, viz [1], lze odvodit následující rovnici pro tlak

$$\frac{\partial\left(\phi(\mathbf{x})p(\mathbf{x},t)\right)}{\partial t} + \sum_{i=1}^{n_c}\theta_i(t)\nabla\cdot\left(c_i(\mathbf{x},t)\mathbf{v}(\mathbf{x},t)\right) = \sum_{i=1}^{n_c}\theta_i(t)f_i(\mathbf{x},t). \tag{3}$$

Veličina θ_i je spojená s molárním objemem směsi a lze ji vypočítat pomocí stavové rovnice $p = p(c_1,...,c_{n_c},T)$ jako

$$\theta_i = \left(\frac{\partial p}{\partial c_i}\right). \tag{4}$$

2.2 Formulace modelu

Nechť $\Omega\subset\mathbb{R}^d, d\in\mathbb{N}$ je omezená oblast a nechť $I\subset\mathbb{R}$ je časový interval. V $\Omega\times I$ řešíme následující soustavu rovnic pro celkové koncentrace $i-$té složky směsi $c_i = c_i(\mathbf{x},t)$

$$\frac{\partial\left(\phi(\mathbf{x})c_i(\mathbf{x},t)\right)}{\partial t} + \nabla\cdot\left(c_i(\mathbf{x},t)\mathbf{v}(\mathbf{x},t)\right) = f_i(\mathbf{x},t), \quad i = 1,...,n_c, \tag{5}$$

kde $\mathbf{v}$ je dáno (2). Rovnice (2), (3) a (5) propojujeme stavovou rovnicí

$$p = p(c_1,...,c_{n_c},T), \tag{6}$$

a předepisujeme následující počáteční a hraniční podmínky

$$\begin{aligned} &c_i(\mathbf{x},0) = c_i^0(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega, \\ &c_i(\mathbf{x},t) = c_i^D(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega, t > 0 \\ &p(\mathbf{x},t) = p^D(\mathbf{x},t), \quad \forall \mathbf{x} \in \Gamma_p, t \in I, \\ &\mathbf{v}(\mathbf{x},t) \cdot \mathbf{n}(\mathbf{x}) = \mathbf{v}^N(\mathbf{x},t), \quad \forall \mathbf{x} \in \Gamma_v, t \in I, \end{aligned} \tag{7}$$

kde $\mathbf{n}$ je vně orientovaný jednotkový normálový vektor k hranici $\partial\Omega$, $\Gamma_p \cup \Gamma_v = \partial\Omega$ a $\Gamma_p \cap \Gamma_v = \emptyset$. Poznamenejme, že rovnice (6) a (7) jsou svázány podmínkou $p^D = p(c_1^D, ..., c_{n_c}^D, T)$.

3 Numerické řešení

Rovnici pro tlak (3) řešíme smíšenou hybridní verzí metody konečných prvků s vyšším řádem přesnosti. Získané rychlostní pole je použito v transportních rovnicích, které jsou řešeny nespojitou Galerkinovou metodou (DG). V dalším se omezíme na jednu dimenzi, tj. v rovnicích (5) - (7) klademe $d = 1$. Oblast $\Omega \subset \mathbb{R}$ pokryjeme numerickou sítí s počtem elementů n_e.

3.1 Diskretizace rychlostního pole

Celkové rychlostní pole je aproximováno pomocí Raviartova-Thomasova prvku prvního stupně jako

$$\mathbf{v}_K(x,t) = \sum_{j=1}^{3} v_{K,j}(t)\mathbf{w}_{K,j}(x), \tag{8}$$

kde $\mathbf{w}_{K,j}$ je bázová funkce prostoru $RT_1(K)$ a $v_{K,j}$ je příslušný stupeň volnosti. Diskretizací Darcyho zákona (2) získáme předpis pro $v_{K,j}$ jako fukci tlaku. Rovnice (2) vynásobíme bazickou funkcí $\mathbf{w}_{K,m}$ a následnou integrací přes element $K = (a,b)$ a použitím per-partes získáme

$$(p\mathbf{w}_{K,m})|_b - (p\mathbf{w}_{K,m})|_a - \int_K p\frac{\partial \mathbf{w}_{K,m}}{\partial x} + \mu_K \sum_{l=1}^{3} v_{K,l} \int_K \mathbf{w}_{K,m}\mathbf{k}^{-1}\mathbf{w}_{K,l} = 0. \tag{9}$$

Na tomto místě poznamenejme, že v jedné dimenzi platí $E \in \partial K = \{a,b\}$, pro $K = (a,b)$. Tlak na elementu K aproximujeme lineární funkcí ve tvaru

$$p_K(x,t) = \sum_{j=1}^{2} \pi_{K,j}(t)\Phi_{K,j}(x), \tag{10}$$

kde $P_1(K) = \text{span}\{\Phi_1, \Phi_2\}$ je prostor polynomů na elementu K stupně nejvýše jedna. Tlak na hranicích elementu popisujeme konstantní funkcí

$$p(x,t)|_E = \hat{\pi}_{E,1}^K(t) \tag{11}$$

V této práci volíme

$$\Phi_{K,1}(x) = 1, \quad \Phi_{K,2}(x) = \frac{2}{b-a}(x - \frac{b+a}{2}) \quad \text{pro} \quad \forall x \in K = (a,b). \tag{12}$$

Rovnice (9) pak přejde na tvar

$$\sum_{l=1}^{3} \tilde{\alpha}_{m,l}^{K} v_{K,l} = \mu_K^{-1} \left(\sum_{i=1}^{2} \beta_{i,m}^{K} \pi_{K,i} - \sum_{E \in \partial K} \hat{\pi}_{E,1}^{K} \chi_E^{K,m} \right), \tag{13}$$

kde jsme pro $K = (a,b)$ označili

$$\begin{aligned} \tilde{\alpha}_{m,l}^{K} &= \int_K \mathbf{w}_{K,m} \mathbf{k}^{-1} \mathbf{w}_{K,l}, \\ \beta_{i,m}^{K} &= \int_K \Phi_{K,i} \frac{\partial \mathbf{w}_{K,m}}{\partial x}, \\ \chi_E^{K,m} &= \begin{cases} -\mathbf{w}_{K,m}(a), & E = \{a\}, \\ \mathbf{w}_{K,m}(b), & E = \{b\}, \end{cases} \end{aligned} \tag{14}$$

Inverzí matice $\tilde{\alpha}^K = [\tilde{\alpha}_{m,l}^K]_{m,l=1}^{3,3}$ získáme explitní vztah pro rychlosti $v_{K,j}$, $j = 1,2,3$

$$v_{K,j} = \mu_K^{-1} \left[\sum_{l=1}^{3} \alpha_{j,l}^{K} \sum_{m=1}^{2} \beta_{l,m} \pi_{K,m} - \sum_{l=1}^{2} \alpha_{j,l}^{K} \sum_{E \in \partial K} \hat{\pi}_{E,1}^{K} \chi_E^{K,j} \right], \quad j = 1,2,3, \tag{15}$$

kde jsme označili $\alpha_{m,l}^K = [\tilde{\alpha}^K]_{m,l}^{-1}$. Protože na veličinu $\mathbf{k}$ je kladena podmínka $\mathbf{k}(x) > 0$, $\forall x \in \Omega$, lze matici $\tilde{\alpha}^K$ invertovat. Podmínku spojitosti normálových komponent rychlostního pole na hranách $E = K \cap K'$ sousedících elementů K, K' zapíšeme pomocí příslušných stupňů volnosti na hraně E. V případě 1D je to pouze jeden stupeň volnosti

$$v_{K,j} + v_{K',j'} = 0, \tag{16}$$

kde index j, resp. j' označuje stupeň volnosti příslušný hraně $E = K \cap K'$ a elementu K, resp. K'. Následující podmínka vyjadřuje rovnost tlaků (resp. spojitost tlaku) na hraně E dvou sousedících elementů K, K'

$$\hat{\pi}_{E,1}^{K} = \hat{\pi}_{E,1}^{K'} =: \hat{\pi}_E. \tag{17}$$

Diskretizací hraničních podmínek (7) získáme

$$\begin{aligned} &\hat{\pi}_E = \hat{\pi}_{E,mean}^{D}, \quad \forall E \subset \Gamma_p, \\ &v_{K,j'} = v_{E,mean}^{N}, \quad \forall E, K : E \subset \Gamma_v, E \in \partial K, \end{aligned} \tag{18}$$

kde $\hat{\pi}_{E,mean}^{D}$ je předepsaná hodnota tlaku p zprůměrovaná na hranu E a $v_{E,mean}^{N}$ je předepsaná rychlost na hraně $E \in \partial K \subset \Gamma_v$, kde index j' opět označuje stupeň volnosti příslušný hraně E elementu K. Symbol $K : E$ označuje elementy K sousedící s hranou

E. Pomocí rovnice (16) lze eliminovat rychlost a získat tak systém lineárních algebraických rovnic pro neznámé π a $\hat{\pi}$

$$\begin{aligned}
&\sum_{K:E\in\partial K}\left[\mu_K^{-1}\sum_{l=1}^{3}\alpha_{j,l}^{K}\sum_{m=1}^{2}\beta_{l,m}\right]\pi_{K,m}(t)-\\
&\sum_{K:E\in\partial K}\left[\mu_K^{-1}\sum_{l=1}^{2}\alpha_{j',l}^{K}\sum_{E'\in\partial K}\chi_{E'}^{K,j'}\right]\hat{\pi}_{E'}^{K}(t)=\sum_{K:E\in\partial K\subset\Gamma_v} v_{E,j}^{N}(t), \qquad \forall E\not\subset\Gamma_p,\\
&\hat{\pi}_E(t)=\hat{\pi}_E^D(t), \qquad \forall E\subset\Gamma_p,
\end{aligned} \tag{19}$$

což lze zapsat v maticové formě

$$\mathbf{R}\cdot\pi-\mathbf{M}\cdot\hat{\pi}=\mathbf{V}. \tag{20}$$

Prvky matic $\mathbf{R},\mathbf{M}$ a vektoru $\mathbf{V}$ lze snadno vyčíst ze soustavy (19).

3.2 Aproximace transportních rovnic a rovnice pro tlak

Transportní rovnice (1) s podmínkami (7) řešíme pomocí diskrétní Galerkinovy metody. Vynásobením (1) bazickou funkcí $\Phi_{K,m}$ a integrací přes element $K=(a,b)$ s použitím per-partes dostaneme pro $i=1,...,n_c$

$$\phi_K\frac{\mathrm{d}}{\mathrm{d}t}\int_K c_i\Phi_{K,m}+[(c_i\Phi_{K,m}\mathbf{v}_K)|_b-(c_i\Phi_{K,m}\mathbf{v}_K)|_a]-\int_K c_i(\mathbf{v}_K\frac{\partial\Phi_{K,m}}{\partial x})=\int_K f_i\Phi_{K,m}, \tag{21}$$

kde jsme označili ϕ_K zprůměrovanou hodnotu porozity ϕ přes element K. Rychlostní pole $\mathbf{v}_K$ aproximujeme pomocí (8) a koncentraci c_i pro $i=1,...n_c$ aproximujeme za použití báze $P_1(K)=\mathrm{span}\{\Phi_{K,1},\Phi_{K,2}\}$ opět pomocí lineární funkce jako

$$c_{i,K}(x,t)=\sum_{j=1}^{2}\xi_{i,K,j}(t)\Phi_{K,j}(x). \tag{22}$$

Jedná se tedy o po elementech lineární aproximaci nespojitou přes hranice elementů. Po úpravě (21) získáme pro $i=1,...,n_c$, $m=1,2$ a každé $K\in T_h$

$$\phi_K\sum_{j=1}^{2}\tilde{\eta}_{m,j}^{K}\frac{\mathrm{d}\xi_{i,K,j}(t)}{\mathrm{d}t}+\sum_{j=1}^{3}\gamma_{m,j}^{i,K}(t)v_{K,j}(t)=F_{i,m}(t), \tag{23}$$

kde jsme označili

$$\begin{aligned}
&\tilde{\eta}_{m,j}^{K}=\int_K\Phi_{K,m}\Phi_{K,j}, \qquad F_{i,m}(t)=\int_K f_i(t)\Phi_{K,m}\\
&\gamma_{m,j}^{i,K}(t)=\sum_{E\in\partial K}\delta_{E,m}^{K,j}\hat{c}_{i,K,E}(t)-\sum_{l=1}^{3}\tau_{j,l}^{K,m}\xi_{i,K,l}(t),\\
&\delta_{E,m}^{K,j}=\begin{cases}-(\mathbf{w}_{K,j}\Phi_{K,m})(a), & E=\{a\},\\ (\mathbf{w}_{K,j}\Phi_{K,m})(b), & E=\{b\},\end{cases} \qquad \tau_{j,l}^{K,m}=\begin{cases}-(\mathbf{w}_{K,j}\Phi_{K,l}\frac{\partial\Phi_{K,m}}{\partial x})(a), & E=\{a\},\\ (\mathbf{w}_{K,j}\Phi_{K,l}\frac{\partial\Phi_{K,m}}{\partial x})(b), & E=\{b\}.\end{cases}
\end{aligned} \tag{24}$$

Ve výrazu (24) pro $\gamma_{m,j}^{i,K}(t)$ vystupuje veličina $\hat{c}_{i,K,E}(t)$, kterou počítáme pomocí upwindového schématu

$$\hat{c}_{i,K,E}(t) = \begin{cases} c_{i,K}(x,t), & v_K n_E \geq 0, \quad x \in E = K \cap K' \\ c_{i,K'}(x,t), & v_K n_E < 0, \quad x \in E = K \cap K' \notin \Gamma_v, \\ c_i^D(x,t), & v_K n_E < 0, \quad x \in E = K \cap K' \in \Gamma_v. \end{cases} \tag{25}$$

Inverzí matice $\tilde{\eta}^K = [\tilde{\eta}_{m,j}^K]_{m,j=1}^{3,3}$ získáme pro $i = 1, ..., n_c$, $m = 1, 2$ a každé $K \in T_h$

$$\frac{\mathrm{d}\xi_{i,K,m}(t)}{\mathrm{d}t} = \frac{1}{\phi_K} \sum_{j=1}^{2} \eta_{m,j}^K F_{i,m}(t) - \frac{1}{\phi_K} \sum_{j=1}^{2} v_{K,j}(t) \left(\sum_{q=1}^{2} \eta_{m,q}^K \gamma_{q,j}^{i,K}(t \right) \equiv \mathcal{C}_m^{i,K}(t). \tag{26}$$

Podobným postupem s rovnicí pro tlak (3) získáme pro $m = 1, 2$ a každé $K \in T_h$

$$\phi_K \sum_{j=1}^{2} \tilde{\eta}_{m,j}^K \frac{\mathrm{d}\pi_{K,j}(t)}{\mathrm{d}t} + \sum_{i=1}^{n_c} \theta_i(t) \sum_{j=1}^{3} \gamma_{m,j}^{i,K}(t) v_{K,j}(t) = \sum_{i=1}^{n_c} \theta_i(t) F_{i,m}(t). \tag{27}$$

Do předchozí rovnice (27) dosadíme vztah (15) pro rychlosti, čímž po úpravách dostaneme

$$\frac{\mathrm{d}\pi_{K,m}(t)}{\mathrm{d}t} = \sum_{j=1}^{2} \sigma_{m,j}^K(t) \pi_K, j(t) + \sum_{j=1}^{2} \lambda_{m,j}^K(t) \hat{\pi}_j^K(t) + \varphi_m^K(t) \equiv \mathcal{P}_m^K(t), \tag{28}$$

kde jsme pomoí inverze $\eta_{m,q}^K = [\tilde{\eta}^K]_{m,q}^{-1}$ označili

$$\begin{aligned} \sigma_{m,j}^K(t) &= -\frac{1}{\phi_K} \mu_K^{-1} \left(\sum_{q=1}^{2} \eta_{m,q}^K \left[\sum_{i=1}^{n_c} \theta_i(t) \sum_{l=1}^{3} \gamma_{q,l}^{i,K}(t) \sum_{r=1}^{3} \alpha_{r,l}^K \beta_{j,r}^K \right] \right), \\ \lambda_{m,j}^K(t) &= \frac{1}{\phi_K} \mu_K^{-1} \left(\sum_{q=1}^{2} \eta_{m,q}^K \left[\sum_{i=1}^{n_c} \theta_i(t) \sum_{l=1}^{3} \gamma_{q,l}^{i,K}(t) \alpha_{j,l}^K \right] \right), \\ \varphi_m^K(t) &= \frac{1}{\phi_K} \left(\sum_{i=1}^{n_c} \theta_i(t) \sum_{q=1}^{2} \eta_{m,q}^K F_{i,q}(t) \right). \end{aligned} \tag{29}$$

3.3 Časová diskretizace

Semi-diskrétní rovnici (28) nyní zdiskretizujeme v čase pomocí $\omega-$schématu Eulerovy metody následovně

$$\pi_{K,m}^{n+1} = \pi_{K,m}^n + \omega \Delta t \mathcal{P}_m^{K,n+1} + (1-\omega) \Delta t \mathcal{P}_m^{K,n}, \quad \forall K \in T_h, m = 1, 2, \tag{30}$$

kde $\omega \in \langle 0, 1 \rangle$. Rovnice (30) přejde pro $\omega = 0$ v explicitní schéma, pro $\omega = 1$ v implicitní schéma a pro $\omega = 1/2$ ve schéma druhého řádu přesnosti Cranka a Nicolsonové. Protože je výraz $\mathcal{P}_m^K$ lineární v neznámé $\pi_{K,m}$, lze dále (30) přepsat na

$$\begin{aligned} \sum_{j=1}^{2} \left[\delta_{m,j} - \omega \Delta t \sigma_{m,j}^{K,n+1} \right] \pi_{K,j}^{n+1} - \sum_{j=1}^{2} \left[(1-\omega) \Delta t \lambda_{m,j}^{K,n+1} \right] \hat{\pi}_{K,j}^{n+1} &= \pi_{K,m}^n + (1-\omega) \Delta t \left(\mathcal{P}_m^{K,n} + \varphi_m^{K,n+1} \right) \\ &= G_m^K, \quad \forall K \in T_h, m = 1, 2. \end{aligned} \tag{31}$$

Předchozí vztah (31) zapíšeme v maticové formě

$$\mathbf{D} \cdot \pi^{n+1} + \mathbf{H} \cdot \hat{\pi}^{n+1} = \mathbf{G}. \tag{32}$$

Rovnici (26) pro koeficienty $\xi_{i,K,m}$ koncentrací $c_{i,K}$ diskretizujeme podobně, viz sekce 3.4.

3.4 Iterační schéma

Pro výpočet numerického řešení jsme využili iterované IMPEC schéma popsané v [3].

1. Pro $n = N_0 : N_T$ opakujeme.
2. Inicializace:
   $l := 0$.
   $\pi_{K,m}^{n+1,0} = \pi_{K,m}^{n}$,
   $\xi_{i,K,m}^{n+1,0} = \xi_{i,K,m}^{n}$.
   hodnoty $\hat{\pi}_{E}^{n+1,0}$ získáme řešením řešení (19),
   hodnoty $v_{K,m}^{n+1,0}$ získáme řešením (15),
   hodnoty $\theta_{i}^{n+1,0}$, získáme řešením (4),
3. Opakuj
   (a) $l := l + 1$.
   (b) Vyřešíme dvojici soustav (20) a (32) pro neznámé $\pi^{n+1,l}$ a $\hat{\pi}^{n+1,l}$

$$\begin{aligned} \mathbf{R} \cdot \pi^{n+1,l} - \mathbf{M} \cdot \hat{\pi}^{n+1,l} &= \mathbf{V}, \\ \mathbf{D} \cdot \pi^{n+1,l} + \mathbf{H} \cdot \hat{\pi}^{n+1,l} &= \mathbf{G}, \end{aligned} \tag{33}$$

   kde lze s výhodou využít toho, že matice D je (blokově) diagonální a tak lze její inverzi snadno vypočítat.
   (c) Vypočteme

$$v_{K,j}^{n+1,l} = \mu_K^{-1} \left[\sum_{l=1}^{3} \alpha_{j,l}^{K} \sum_{i=1}^{2} \beta_{l,i} \pi_{K,i}^{n+1,l} - \sum_{l=1}^{2} \alpha_{j,l}^{K} \sum_{E \in \partial K} \hat{\pi}_{E}^{K,n+1,l} \chi_{E}^{K,m} \right], \quad j = 1, 2, 3. \tag{34}$$

   (d) Nalezneme neznámé hodnoty $\xi_i^{n+1,l}$ pro $i = 1, ...n_c$, diskretizací (26) z rovnice

$$\begin{aligned} \xi_{i,K,m}^{n+1,l} &= \xi_{K,m}^{n} + \omega \Delta t \tilde{\mathcal{C}}_{m}^{i,K,n+1,l,l-1} + (1-\omega) \Delta t \mathcal{C}_{m}^{i,K,n}, \quad \forall K \in T_h, m = 1, 2, \\ \tilde{\mathcal{C}}_{m}^{i,K,n+1,l,l-1} &= \frac{1}{\phi_K} \sum_{j=1}^{2} \eta_{m,j}^{K} F_{i,m}^{n+1} - \frac{1}{\phi_K} \sum_{j=1}^{3} v_{K,j}^{n+1,l} \left(\sum_{q=1}^{2} \eta_{m,q}^{K} \gamma_{m,j}^{i,K,n+1,l-1} \right). \end{aligned} \tag{35}$$

   (e) Iterace končí, pokud platí

$$\max\{ \frac{||p^{n+1,l} - p^{n+1,l-1}||_{L_2}^2}{||p^{n+1,l}||_{L_2}^2}, \sum_{i=1}^{n_c} \frac{||c_i^{n+1,l} - c_i^{n+1,l-1}||_{L_2}^2}{||c_i^{n+1,l}||_{L_2}^2}, \sum_{i=1}^{n_c} \frac{||\theta_i^{n+1,l} - \theta_i^{n+1,l-1}||_{L_2}^2}{||\theta_i^{n+1,l}||_{L_2}^2} \} < \text{tol}, \tag{36}$$

kde tol je předepsaná přesnost a

$$p_K = \sum_{j=1}^{2} \pi_K, j\Phi_{K,j}, \quad c_{i,K} = \sum_{j=1}^{2} \xi_{i,K,j}\Phi_{K,j}, \quad i = 1, ...n_c. \tag{37}$$

Pokud jsme dosáhli požadované přesnosti, pokládáme $\pi^{n+1} = \pi^{n+1,l}$, $v^{n+1} = v^{n+1,l}$ a $\xi_i^{n+1} = \xi_i^{n+1,l}$, $i = 1, ...n_c$, $n := n + 1$ a jdeme na krok 1. V opačném případě jdeme na krok (a).

4 Výsledky

Navržené numerické schéma otestujeme na 1D úloze, jejíž analytické řešení je známo. Uvažujeme úlohu pro $\Omega = \langle a, b\rangle$, $I = (t_0, \tau)$, $n_c = 1$, $f_1(x,t) = 0$, $\phi = 1$, $\mathbf{k} = 1$, $\mu = 1/2$ se stavovou rovnicí $p = c$. Řešíme tedy úlohu

$$\frac{\partial c(x,t)}{\partial t} - \frac{\partial}{\partial x}\left(c(x,t)\mathbf{v}(x,t)\right) = 0, \quad \mathbf{v} = -2\frac{\partial c}{\partial x}, \tag{38}$$

s počáteční a hraničními podmínkami

$$\begin{aligned} &c(x,0) = B(x,t_0), && \forall x \in \Omega, \\ &p(x,t) = B(x,t), && \forall x \in \Gamma_p, t \in (t_0,\tau), \\ &\mathbf{v}(x,t) = 0, && \forall x \in \Gamma_v, t \in (t_0,\tau), \end{aligned} \tag{39}$$

kde $B = B(x,t)$ je Barenblattovo řešení úlohy ve tvaru

$$B(x,t) = t^{-\frac{1}{3}}\left(1 - \frac{x^2}{12}t^{-\frac{2}{3}}\right)_+, \tag{40}$$

kde $(z)_+ = \max\{z, 0\}$. Dále volíme oblast $\Omega = \langle 0, 100\rangle$, počáteční čas $t_0 = 4 \times 10^3$ a koncový čas $\tau = 10^6$. Numerické řešení je srovnáno s analytickým (40) pomocí experimentálního řádu konvergence (EOC) za použití prvků RT_0 a RT_1. Analytické řešení je projektováno na numerickou síť, kde je počítáno numerické řešení a to vždy do krajů intervalu elementu. Analytické řešení je tak aproximováno po elementech lineární funkcí. Chybu E_n mezi analytickým a numerickým řešením počítáme ve třech normách L_1, L_2 a L_∞. Časový krok volíme $\Delta t = 400$ a časovou diskretizaci volíme schéma Cranka-Nicolsonové, tj. $\omega = 0.5$ v (31) a (35).

Z tabulky 3 vidíme, že zjemňujeme-li časový krok úměrně prostorovému kroku $\Delta t \sim 1/n_e$, dochází ke snížení řádu přesnosti, i když je použita metoda druhého řádu Cranka-Nicolsonové. Problém patrně spočívá v kroku (d) iteračního algoritmu, kdy koncentrace na nové časové hladině počítáme explicitně. Ačkoliv se toto explicitní iterační schéma přibližuje k implicitnímu schématu pomocí vnitřního iteračního cyklu, schéma bude mít patrně zlomkový řád přesnosti v čase.

Tabulka 1: Chyby koncentrací c v čase $\tau = 10^6$ vůči analytickému řešení a EOC pro RT_0. Pro časový krok platí $\Delta t \sim 1/n_e^2$.

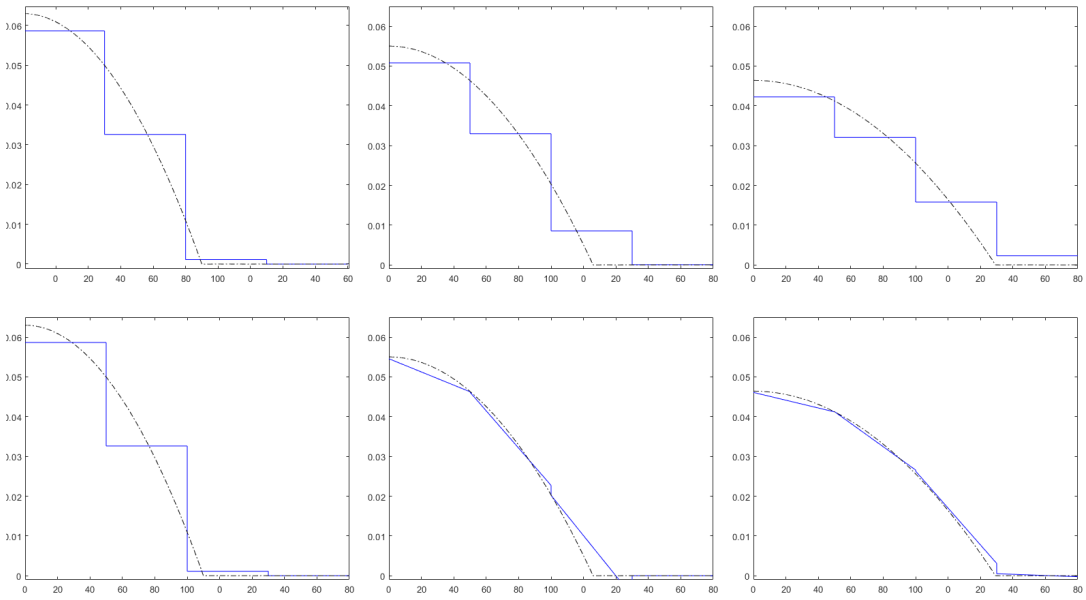
Síť (n_e)	Δt	$\|\|E_n\|\|_{L_1}$	EOC_1	$\|\|E_n\|\|_{L_2}$	EOC_2	$\|\|E_n\|\|_{L_\infty}$	EOC_∞
4	100	5.6314×10^{-3}	–	7.1172×10^{-4}	–	1.8286×10^{-4}	–
8	25	2.8488×10^{-3}	0.9838	3.5985×10^{-4}	0.9839	9.7048×10^{-5}	0.9131
16	6.25	1.4332×10^{-3}	0.9911	1.8096×10^{-4}	0.9917	5.0018×10^{-5}	0.9563
32	1.5625	7.1949×10^{-4}	0.9942	9.0770×10^{-5}	0.9954	2.5397×10^{-5}	0.9778
64	0.39063	3.6063×10^{-4}	0.9965	4.5467×10^{-5}	0.9974	1.2797×10^{-5}	0.9888
128	0.097656	1.8057×10^{-4}	0.9979	2.2757×10^{-5}	0.99853	6.4240×10^{-6}	0.99432

Tabulka 2: Chyby koncentrací c v čase $\tau = 10^6$ vůči analytickému řešení a EOC pro RT_1. Pro časový krok platí $\Delta t \sim 1/n_e^2$.

Síť (n_e)	Δt	$\|\|E_n\|\|_{L_1}$	EOC_1	$\|\|E_n\|\|_{L_2}$	EOC_2	$\|\|E_n\|\|_{L_\infty}$	EOC_∞
4	100	3.4048×10^{-4}	–	3.8893×10^{-5}	–	8.3915×10^{-6}	–
8	25	8.5181×10^{-5}	1.9990	9.7260×10^{-6}	1.9996	2.1038×10^{-6}	1.9959
16	6.25	2.1300×10^{-5}	1.9997	2.4317×10^{-6}	1.9999	5.2663×10^{-7}	1.9981
32	1.5625	5.3252×10^{-6}	2.0000	6.0794×10^{-7}	2.0000	1.3174×10^{-7}	1.9991
64	0.39063	1.3300×10^{-6}	2.0008	1.5195×10^{-7}	2.0003	3.2944×10^{-8}	1.9996
128	0.097656	3.3266×10^{-7}	1.9998	3.7990×10^{-8}	2.0000	8.2373×10^{-9}	1.9998

Tabulka 3: Chyby koncentrací c v čase $\tau = 10^6$ vůči analytickému řešení a EOC pro RT_1. Pro časový krok platí $\Delta t \sim 1/n_e$.

Síť (n_e)	Δt	$\|\|E_n\|\|_{L_1}$	EOC_1	$\|\|E_n\|\|_{L_2}$	EOC_2	$\|\|E_n\|\|_{L_\infty}$	EOC_∞
64	12.5	1.3061×10^{-6}	–	1.5175×10^{-7}	–	3.3731×10^{-8}	–
128	6.25	3.2024×10^{-7}	2.0281	3.8283×10^{-8}	1.9869	8.9815×10^{-9}	1.9090
256	3.125	8.0228×10^{-8}	1.9970	1.0033×10^{-8}	1.9319	2.5192×10^{-9}	1.8340
512	1.5625	2.2174×10^{-8}	1.8553	2.9758×10^{-9}	1.7535	7.6625×10^{-10}	1.7171
1024	0.78125	8.6899×10^{-9}	1.3514	1.0664×10^{-9}	1.4805	2.5281×10^{-10}	1.5997
2048	0.39063	2.4776×10^{-9}	1.8104	$2.9762e \times 10^{-10}$	1.8413	6.8590×10^{-11}	1.8820



Obrázek 1: Kvalitativní porovnání numerického řešení v několika časových hladinách za použití RT_0 (nahoře) a RT_1 (dole). Čerchovanou čárou je vyzobrazeno analytické řešení a plnou čárou numerické řešení.

5 Závěr

V této práci jsme implementovali iterační metodu fully mass-conservative iterative IMPEC společně s MHFEM a DG vyšších řádů přesnosti pro modelování jedno-fázového proudění vícesložkové směsi v porézním prostředí. Pomocí srovnání se známým analytickým řešením jsme ověřili vyšší řád přesnosti v prostoru. Při použití schématu Cranka-Nicolsonové s druhým řádem přesnosti v čase jsme narazili na jev, kdy se chyba časové diskretizace nesnižuje s řádem dva ale stále více než jedna. To spolu s implementací metody ve dvou dimenzích bude předmětem dalšího zkoumání.

Literatura

[1] G. Ács, S. Doleschall, É. Farkas. *General purpose compositional model.* Society of Petroleum Engineers Journal, Vol.: 25, Issue: 4 (1985) 543–553

[2] Huangxin Chen, Xiaolin Fan, Shuyu Sun. *A Fully Mass-Conservative Iterative IMPEC Method for Multicomponent Compressible Flow in Porous Media.* Journal of Computational and Applied Mathematics, Vol.: 362 (2019) 1-21

[3] O. Polívka, J. Mikyška. *Combined Mixed-Hybrid Finite Element–Finite Volume Scheme for Computation of Multicomponent Compressible Flow in Porous Media.* Numerical Mathematics and Advanced Applications 2011, Springer-Verlag, Berlin, Heidelberg (2013) 559-567

[4] D. Boffi, F.Brezzi, M. Fortin. *Mixed Finite Element Methods and Applications.* Springer Series in Computational Mathematics, Vol.: 44, Springer Science & Business Media, (2013)

Fusion of Probabilistic Knowledge in Multi-Agent Decision Making*

František Hůla

2nd year of PGS, email: hulafra1@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Tatiana Valentine Guy, Department of Adaptive Systems
Institute of Information Theory and Automation, CAS

Miroslav Kárný, Department of Adaptive Systems
Institute of Information Theory and Automation, CAS

Abstract. Bayesian decision maker accumulates its knowledge into probability distribution of an unknown parameter within a parametric model serving for prediction of future observations. Bayes' rule is the unique accumulation mechanism whenever realisations of relevant data are observed. In multiple-participant decision making, an independent knowledge is often compressed by another participant into the probability distribution describing the same future observations. The use of this probability distribution may significantly spare the effort spent by the knowledge-accumulating participant. Particular recommendations how to exploit this knowledge description exist, but a systematic methodology is missing. This paper formulates and solves a widely applicable knowledge processing respecting the information fusion principle.

Keywords: decision-making, information fusion, cooperation

Abstrakt. Bayesovský rozhodovač ukládá svoji znalost jako pravděpodobnostní rozdělení nad neznámým parametrem skrze parametrický model sloužící pro predikci budoucích pozorování. Bayesovo pravidlo je jedinečný mechanismus akumulace nabyté znalosti, kdykoliv je pozorována realizace relevantních dat. V multiagentním rozhodování je nezávisle na sobě získaná znalost často shromažďována jiným agentem do pravděpodobnostní distribuce popisující ta samá budoucí pozorování. Použití této pravděpodobnostní distribuce může podstatně ušetřit úsilí vynaložené agentem shromažďujícím potřebnou znalost o parametrech. Jistá doporučení jak využít popis této znalosti existují, nicméně systematická metodologie chybí. Tato práce formuluje a řeší široce aplikovatelné zpracování znalosti s ohledem na princip fúze informace.

Klíčová slova: rozhodování, fúze informace, spolupráce

Full paper: F. Hůla, M. Kárný, T.V. Guy. *Fusion of Accumulated Probabilistic Knowledge in Multiple-Participant Decision Making.* Submitted to Information Fusion.

*This work has been supported by LTC18075

Phase Field Model of Phase Transitions at Microscale*

Jakub Kantner

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

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

Abstract. This contribution studies the phase transitions by following the time development of the interface between a solid and liquid phase in a medium. Phase transitions can been found both in the nature and in the industry. In the first part, the used mathematical model is described. In the second part, a numerical solution based on the finite-difference method is given. Finally in the last part of the contribution, the simulations verifying the used model are shown as well as some preliminary results of an anisotropic crystal growth.

Keywords: phase field, phase transition, crystal growth, dendrites

Abstrakt. V tomto příspěvku studujeme fázové přechody na základě vývoje rozhraní mezi pevnou a kapalnou fází v médiu. Fázové přechody se vyskytují jak v přírodě, tak i v průmyslu. V první části příspěvku je popsán použitý matematický model. V druhé části je ukázáno použité numerické řešení založené na metodě konečných diferencí. Nakonec jsou v poslední části příspěvku ukázány simulace potvrzující správnost pouzitého modelu, jakož i předběžné výsledky anizotropního růstu krystalů.

Klíčová slova: fázové pole, fázové přechody, růst krystalů, dendrity

1 Introduction

The main motivation of this paper is to study and model a growth of crystals. The crystallisation is a phenomenon occurring both in the nature (e.g. snowflakes) and in the industry (from semi-conductors to jet engine turbines).

The study of the crystallization is limited to the time development of the interface between a solid and liquid phase. Starting from a nucleus of solid phase the liquid phase can either solidify at a nucleus boundary – enlarging the crystal – or the crystal itself can melt and turn back to the liquid phase. This process can also be anisotropic – the crystal grows faster in certain direction than in others creating dendrites.

This work follows papers by M. Beneš, which study the microstructure growth during solidification [1], [2], [3] and papers by A. Žák, which study the growth in porous media [8], [9]. This article is divided into two sections. In the first section, the used mathematical model is described and a numerical scheme is presented. In the second section, the

*This work has been partially supported by the project "Investigation of shallow subsurface flow with phase transitions" No. 17-06759S of Czech Science Foundation.

numerical scheme is first verified by two settings with known analytical solutions. Then several interesting cases are shown.

2 Mathematical model

In this section, we present the model of phase transitions and the numerical scheme used to obtain a numerical solution.

A model, taken from [1], was used to describe the behaviour of the phase transitions. This model consists of two partial differential equations for two variables p and u representing the phase and the temperature, respectively. It takes form:

$$\begin{aligned} \alpha\xi^2\partial_t p &= g(\Theta)\xi^2\Delta p + f_0(p) + F(u,p) && \text{in } (0,T)\times\Omega, \\ \rho c\partial_t u &= \lambda\Delta u + L\partial_t p && \text{in } (0,T)\times\Omega, \\ u|_{t=0} &= u_{ini}, \quad v|_{t=0} = v_{ini} && \text{in } \Omega, \\ u|_{\partial\Omega} &= \gamma_u, \quad v|_{\partial\Omega} = \gamma_v && \text{on } (0,T)\times\partial\Omega. \end{aligned} \tag{1}$$

In this paper the right hand side function f_0 always takes the form

$$f_0(p) = ap(1-p)\left(p - \frac{1}{2}\right), \tag{2}$$

where a is a constant, while the right hand side source term $F(u,p)$ is used in two different forms. The first one, denoted as *Model 1*, is

$$F(u,p) = -b\beta\xi^2(u-u^*) \tag{3}$$

and the second one, denoted as *Model 3*, is

$$F(u,p) = -\beta\xi^2|\nabla p|(u-u^*), \tag{4}$$

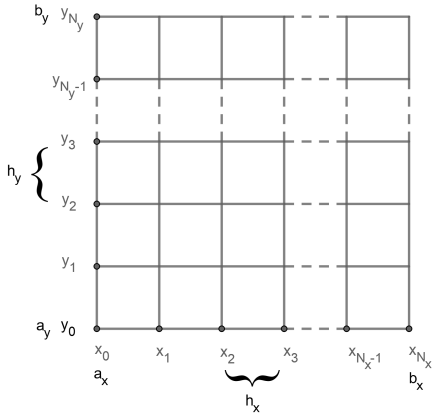
where $|\cdot|$ is a norm. The function $g = g(\Theta)$, which determines the anisotropy, is set to a constant value 1 in most experiments in this paper. The only exceptions are experiments with dendrites in Sections 3.2.2 and 3.2.3, where $g(\Theta)$ of the following form is used

$$g(\Theta) = 1 + A(1-m^2)\cos(m(\Theta-\Theta_0)). \tag{5}$$

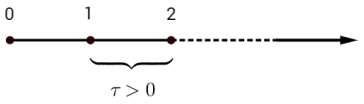
Variable Θ is dependent on the gradient of p as $\Theta = \arctan\left(\frac{\partial_x p}{\partial_y p}\right)$. The values of parameters Θ_0 and m determine the anisotropy type. In nature, mostly 4-fold (for $m=4$) and 6-fold (for $m=6$) symmetries occur. Furthermore, value of parameter A has to satisfy the following condition

$$A < \frac{1}{m^2-1}. \tag{6}$$

The physical properties used are L the latent heat, u^* the transition temperature (or melting point), λ the heat conductivity, ρ the density and c the heat capacity. λ, ρ and c are in general functions of u, but in this paper we assume them to be constant. Parameters α, β, ξ and b are also constant.



(a) Spatial mesh



(b) Mesh for discretization in time

Figure 1: 1a: A discretization of a rectangle $(a_x, b_x) \times (a_y, b_y)$ into a mesh with $(N_x + 1) \times (N_y + 1)$ nodes with equidistant spatial steps h_x and h_y. 1b: A discretization of time axis with a time step τ and the initial time level $t_0 = 0$. [5]

2.1 Numerical Solution

This subsection describes the discretization of space and time and the numerical scheme used to obtain a numerical solution.

2.1.1 Discretization in Space and Time

The numerical solution is calculated on a discrete mesh and in discrete time steps. The mesh is created on a rectangular domain $\Omega = (a_x, b_x) \times (a_y, b_y)$ as a set of nodes

$$\{(ih_x, jh_y) | i = 0, \ldots, N_x - 1, j = 0, \ldots, N_y - 1\} \tag{7}$$

where N_x, $N_y \in \mathbb{N}$ are numbers of mesh nodes in x and y axis, respectively, and $h_x = \frac{b_x - a_x}{N_x}$ and $h_y = \frac{b_y - a_y}{N_y}$ are the lengths of the spatial steps. The mesh is in Figure 1a. The coordinates of $(i, j)-$th point can be obtained as $x_i = a_x + ih_x$ and $y_j = a_y + jh_y$. The discretization of time is obtained by division of time into discrete time steps of length τ. This is shown in Figure 1b. Starting from an initial time $t_0 = 0$ the k-th time level is $t_k = k\tau$.

2.1.2 Finite-Difference Scheme

The functions u, p in (1) depend on time t and spatial coordinates x, y, i.e., $u = u(t, x, y), p = p(t, x, y)$. Let the notation of t_k, x_i a y_j be as stated above. Furthermore, we denote

$$\begin{aligned} u(t_k, x_i, y_j) &= u_{i,j}^k, \\ p(t_k, x_i, y_j) &= p_{i,j}^k. \end{aligned} \tag{8}$$

In what follows, w substitutes the variables u and p.

This scheme uses the forward difference to substitute the first time derivative $\partial_t w$ as

$$\partial_t w_{i,j}^k = \frac{1}{\tau}\left(w_{i,j}^{k+1} - w_{i,j}^k\right) + \mathcal{O}(\tau) \tag{9}$$

and the central difference to substitute the second spatial derivative Δw as

$$\begin{aligned}\Delta w_{i,j}^k &= \partial_{xx} w_{i,j}^k + \partial_{yy} w_{i,j}^k \\ &= \frac{1}{h_x^2}\left(w_{i+1,j}^k - 2w_{i,j}^k + w_{i-1,j}^k\right) + \frac{1}{h_y^2}\left(w_{i,j+1}^k - 2w_{i,j}^k + w_{i,j-1}^k\right) + \mathcal{O}(h_x^2 + h_y^2).\end{aligned} \tag{10}$$

The gradient $|\nabla p|$ in Model 3 is approximated as follows

$$|\nabla p|_{i,j}^k = \sqrt{\frac{1}{2}|\nabla p_{i,j}^k|^2 + \frac{1}{2}|\overline{\nabla} p_{i,j}^k|^2}, \tag{11}$$

where $|\cdot|$ is the standard norm in $\mathbb{R}^2$ and

$$\begin{aligned}\nabla p_{i,j}^k &= \begin{pmatrix}\partial_x p_{i,j}^k \\ \partial_y p_{i,j}^k\end{pmatrix} = \begin{pmatrix}\frac{p_{i+1,j}^k - p_{i,j}^k}{h_1} \\ \frac{p_{i,j+1}^k - p_{i,j}^k}{h_2}\end{pmatrix}, \\ \overline{\nabla} p_{i,j}^k &= \begin{pmatrix}\partial_{\overline{x}} p_{i,j}^k \\ \partial_{\overline{y}} p_{i,j}^k\end{pmatrix} = \begin{pmatrix}\frac{p_{i,j}^k - p_{i-1,j}^k}{h_1} \\ \frac{p_{i,j}^k - p_{i,j-1}^k}{h_2}\end{pmatrix}\end{aligned} \tag{12}$$

are the forward and backward gradients of p.

Finally, we set $\Theta_{i,j}^k$ to be equal to

$$\Theta_{i,j}^k = \arctan\left(\frac{h_y}{h_x} \cdot \frac{p_{i+1,j}^k - p_{i-1,j}^k}{p_{i,j+1}^k - p_{i,j-1}^k}\right). \tag{13}$$

Applying (8)–(13) to Problem (1) we obtain the final numerical scheme as follows

$$\begin{aligned}\alpha\xi^2 \frac{p_{i,j}^{k+1} - p_{i,j}^k}{\tau} &= g(\Theta_{i,j}^k)\xi^2\left(\frac{p_{i+1,j}^k - 2p_{i,j}^k + p_{i-1,j}^k}{h_x^2} + \frac{p_{i,j+1}^k - 2p_{i,j}^k + p_{i,j-1}^k}{h_y^2}\right) + f_0(p_{i,j}^k) + F(u_{i,j}^k, p_{i,j}^k), \\ \rho c \frac{u_{i,j}^{k+1} - u_{i,j}^k}{\tau} &= \lambda\left(\frac{u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k}{h_x^2} + \frac{u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k}{h_y^2}\right) + L\frac{p_{i,j}^{k+1} - p_{i,j}^k}{\tau},\end{aligned} \tag{14}$$

where $f_0(p_{i,j}^k)$ and $F(u_{i,j}^k, p_{i,j}^k)$ are obtained from (2) and (3)-(4) based on the selected model by substitution of u and p for $u_{i,j}^k$ and $_{i,j}^k$ from (8). $|\nabla p_{i,j}^k|$ is then substituted using (11). We can notice, that only one term in (14) in each equation is at time level t_{k+1}. By expressing this term we can calculate new values of u and p in time t_{k+1} using terms at time t_k.

3 Computational studies

In this section, various conducted computations are presented. In the first subsection, computations verifying the model and the scheme are conducted and in the second subsection some preliminary results of a dendritic growth are shown.

parameter	R_0	α	ξ	ρ	c	λ	a	b	β	u^*	L
value	0.15	10	0.01	1	1	0.1	1	0.118	100	1	0

Table 1: Values of parameters used in the Level Set Verification experiment.

Parameter	**Notation**	**Value**
domain	Ω	$(0,1)\times(0,1)$
x-axis spatial step	h_x	0.01
y-axis spatial step	h_y	0.01
time step	τ	$h_x \cdot h_y = 10^{-4}$
time of computation	T	1.9

Table 2: Numerical parameters in the Level Set Verification experiment.

3.1 Model verification

To verify the numerical solution we use two examples in which it can be compared with a known analytical solution.

3.1.1 Level Set Verification

In this experiment, we set the initial and boundary conditions as follows

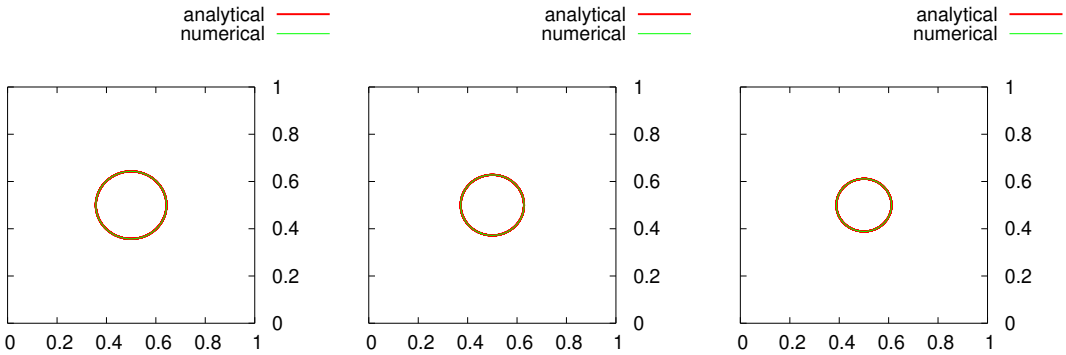
$$\begin{aligned} u|_{t=0} = 1, \quad p|_{t=0} &= \frac{1}{2}\left(1-\tanh\left(\frac{8|x|}{2R_0}-4\right)\right) && \text{in } \Omega, \\ u|_{\partial\Omega} = 1, \quad p|_{\partial\Omega} &= \frac{1}{2}\left(1-\tanh\left(\frac{8|x|}{2R_0}-4\right)\right) && \text{on } (0,T)\times\partial\Omega, \end{aligned} \tag{15}$$

where R_0 is a positive parameter and $|x|$ denotes the $\mathbb{R}^2$ norm of $x \in \Omega$. The value of R_0 as well as the other values of parameters of (1) are in Table 1. Numerical characteristics can then be found in Table 2. We then follow the time development of the intersection of the phase field p with a constant surface with value equal to 0.5. This intersection has a circular shape whose radius $R = R(t)$ develops in accordance with the analytical solution of the problem

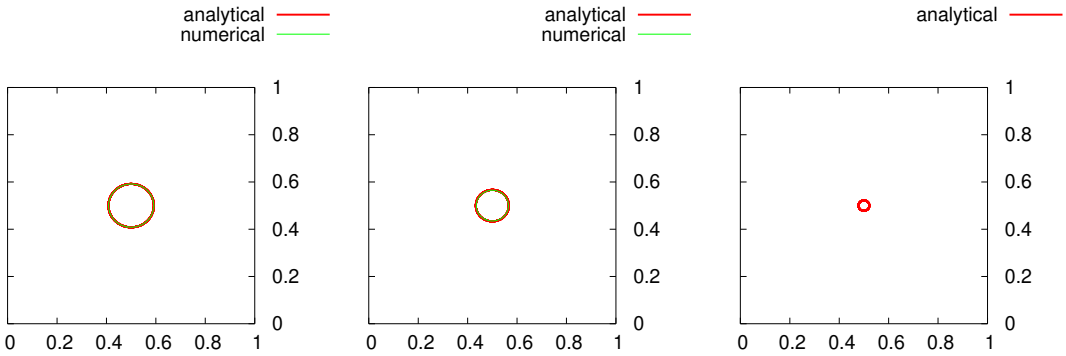
$$R(t) = \sqrt{R_0^2 - \frac{2t}{\alpha}}, \tag{16}$$

where values of parameters R_0 and α are identical with those in Table 1.

The comparison of numerically obtained circular intersection and analytical circle with radius $R = R(t)$ given by (16) can be seen in Figure 2. Numerical solution is shown by the thin green curve while the analytical one by the thick red curve. We can see that the numerical solution closely approximates the analytical one.



(a) p in time $t = 0$ (b) p in time $t = 0.03$ (c) p in time $t = 0.05$



(d) p in time $t = 0.07$ (e) p in time $t = 0.09$ (f) p in time $t = 0.11$

Figure 2: The time development of the Level Set Verification experiment. The red circle represents the analytical solution and the green circle represents the numerical solution.

parameter	R_0	α	ξ	ρ	c	λ	λ_0	a	β	u^*	L	ε
value	0.1	20	h_x	1	1	0.15	0.5	2	20	0	1	0.02

Table 3: Values of parameters used in the Volcano experiment. Note that value of ξ is equal to spatial step of the mesh that can be found in Table 4.

3.1.2 Volcano

This subsection provides another verification of the numerical scheme. In [6] an analytical solution for the temperature variable u of Problem (1) is given in the following form

$$\tilde{u}(x,t) = \begin{cases} U(t) & x \in \Omega \subset \mathbb{R}^2, |x| \leq R(t), \\ U(t) + T\left(\frac{|x|}{R(t)}\right) & x \in \Omega \subset \mathbb{R}^2, |x| > R(t)., \end{cases} \tag{17}$$

where

$$\begin{aligned} R(t) &= \sqrt{R_0^2 + 2\lambda_0 t}, \quad U(t) = \frac{-\varepsilon(\lambda_0+1)}{R(t)}, \\ T(s) &= -\lambda_0 e^{\lambda_0/2} \int_1^s \frac{e^{-(\lambda_0/2)z^2}}{z} \mathrm{d}z. \end{aligned} \tag{18}$$

If we set initial and boundary conditions to

$$\begin{aligned} u|_{t=0} &= \tilde{u}(x,0), \quad p|_{t=0} = \tilde{p}(x), \\ u|_{\partial\Omega} &= \tilde{u}(x,t), \quad p|_{\partial\Omega} = 0, \end{aligned} \tag{19}$$

where $\tilde{p}$ is defined as

$$\tilde{p}(x) = \begin{cases} 1 & x \in \Omega \subset \mathbb{R}^2, |x| \leq R_0 - 2\xi, \\ 0 & x \in \Omega \subset \mathbb{R}^2, |x| > R_0 + 2\xi, \\ |x| + \frac{R_0+2\xi}{4\xi} & \text{otherwise}, \end{cases} \tag{20}$$

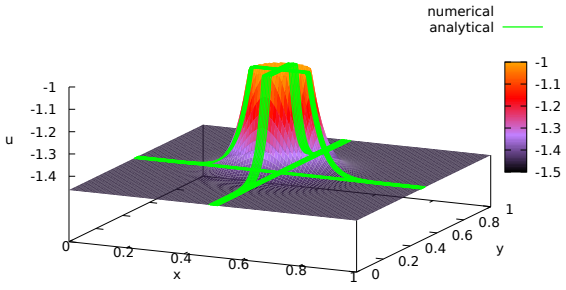
the numerical solution of Problem (1) for temperature u at time t should correspond with analytical solution $\tilde{u}(x,t)$ from (17) in the whole domain Ω. Figure 3 shows that the numerical solution (multicoloured surface) is in accordance with the analytical solution (two bands of this solution are shown in green color). The parameter values are in Table 3. The numerical parameters are then in Table 4.

3.2 Dendritic Patterns

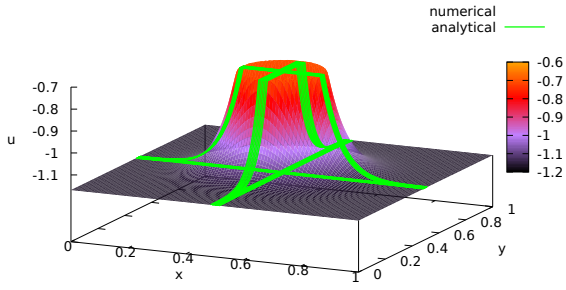
This subsection contains preliminary numerical results of dendritic growth of a crystal. All three examples shows growth of a crystal from a round nucleus. The first computation shows an isotropic stable pattern growth, while others show development of instabilities into dendrites – namely four or six – in an anisotropic medium. In all experiments the Model 3 right hand side function $F(u,p)$ (4) is used.

Parameter	Notation	Value
domain	Ω	$(0,3) \times (0,3)$
x-axis spatial step	h_x	$2.5 \cdot 10^{-3}$
y-axis spatial step	h_y	$2.5 \cdot 10^{-3}$
time step	τ	$0.01 \cdot h_x \cdot h_y = 6.25 \cdot 10^{-8}$
time of computation	T	0.1

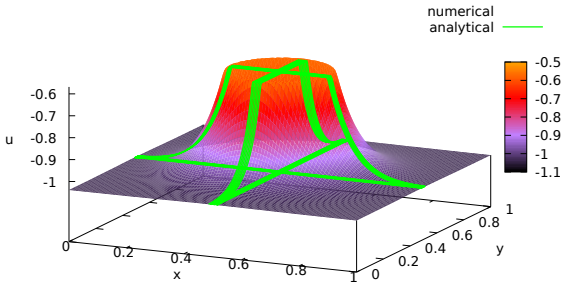
Table 4: Numerical parameters in the Volcano experiment.



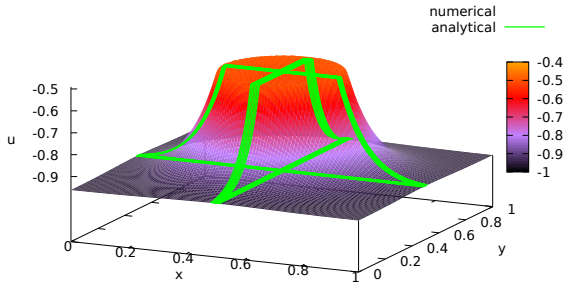
(a) u in time $t = 0$



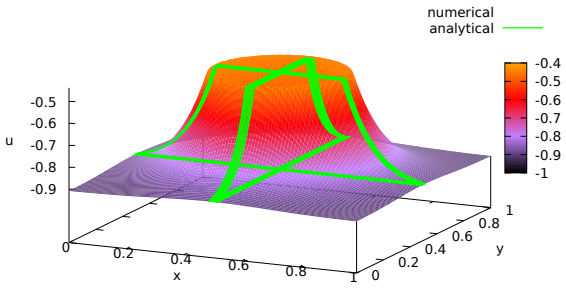
(b) u in time $t = 0.005$



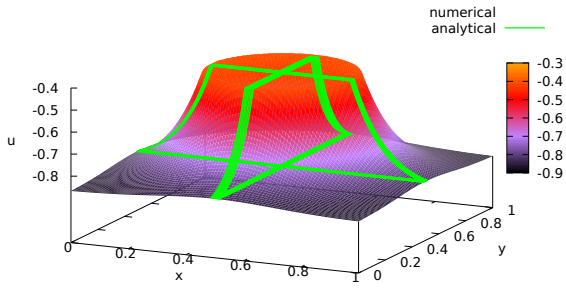
(c) u in time $t = 0.01$



(d) u in time $t = 0.015$

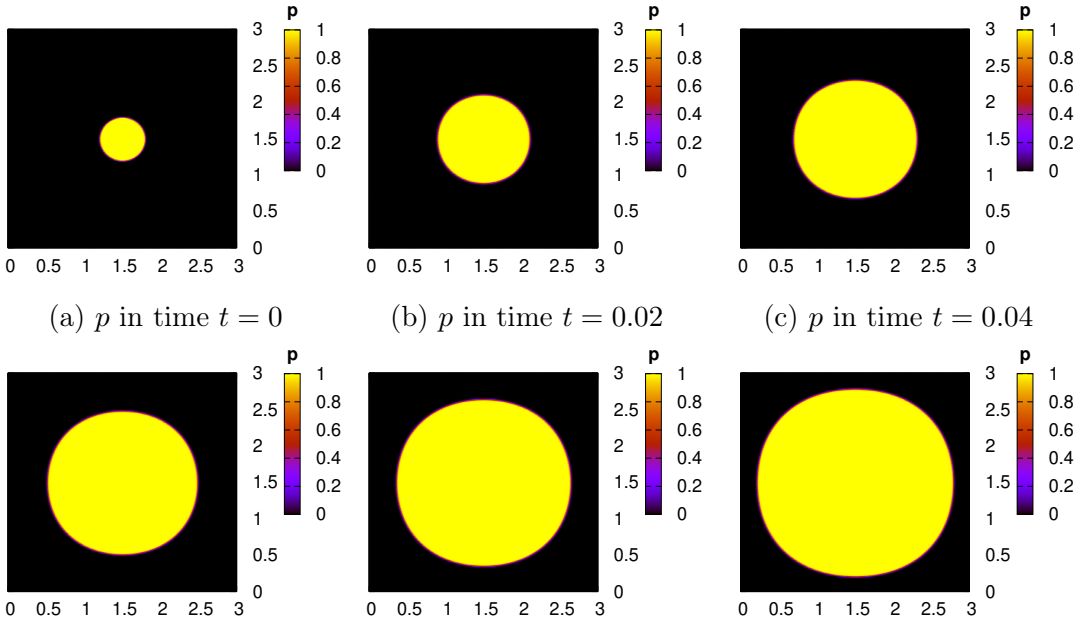


(e) u in time $t = 0.02$



(f) u in time $t = 0.025$

Figure 3: The time development of the Volcano experiment. The coloured surface shows the numerically obtained temperature u while the green bands show a part of the analytical solution.



(a) p in time $t = 0$ (b) p in time $t = 0.02$ (c) p in time $t = 0.04$

(d) p in time $t = 0.06$ (e) p in time $t = 0.08$ (f) p in time $t = 0.1$

Figure 4: The time development of a crystal growth without dendrites.

No. \ Parameters	R_0	α	ξ	ρ	c	λ	a	β	u^*	L	A	m	Θ_0
no dendrites	0.3	1	$4 \cdot 10^{-3}$	1	1	1	2	50	0	1	-	-	-
four dendrites	0.3	1	$4 \cdot 10^{-3}$	1	1	1	2	200	0	1	0.06	4	0.
six dendrites	0.3	1	$4 \cdot 10^{-3}$	1	1	1	2	200	0	1	0.02	6	0.2

Table 5: Values of parameters used in the experiments with no dendrites, four dendrites and six dendrites.

3.2.1 Growth without Dendrites

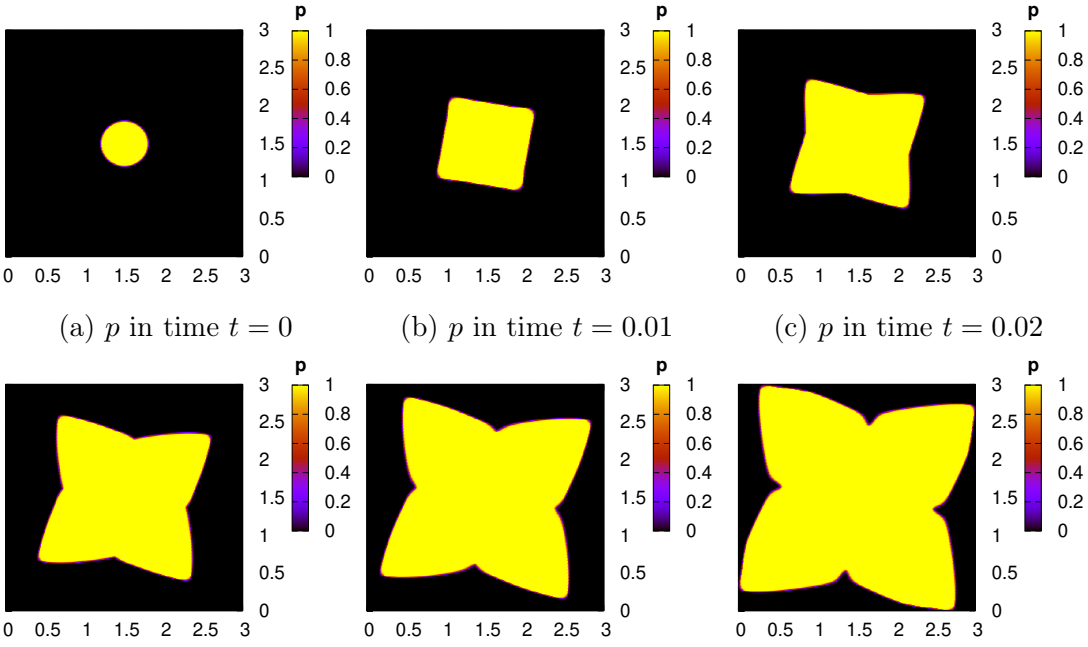
In the first experiment, we study a growth of a crystal from a round nucleus with the radius R_0 in an isotropic medium. The time development in Figure 4 shows only a growing circle which eventually fills the whole domain Ω. No dendrites appear in this case. Their growth is blocked not only by the isotropic medium but also by a relatively low value of parameter β. The value of this parameter can be found in the first row of Table 5 together with the rest of parameter values used in this experiment. Table 6 then shows the used numerical parameters.

3.2.2 Growth with Four Dendrites

In this experiment the anisotropy function $g(\Theta)$ is used for the first time in this paper. The initial settings are the same as in the previous experiment. Moreover, most of the parameter values are the same as in the previous experiment except the value β. Since the anisotropy is present in the model, we also have to add three new parameters A, m and Θ_0. The parameter values are summarized in the second row of Table 5. In Figure

Parameter	Notation	Value
domain	Ω	$(0,3) \times (0,3)$
x-axis spatial step	h_x	$5 \cdot 10^{-3}$
y-axis spatial step	h_y	$5 \cdot 10^{-3}$
time step	τ	$0.01 \times h_x \cdot h_y = 2.5 \cdot 10^{-7}$
time of computation	T	0.1

Table 6: Numerical parameters in the experiment of crystal growth without dendrites.



(a) p in time $t = 0$ (b) p in time $t = 0.01$ (c) p in time $t = 0.02$

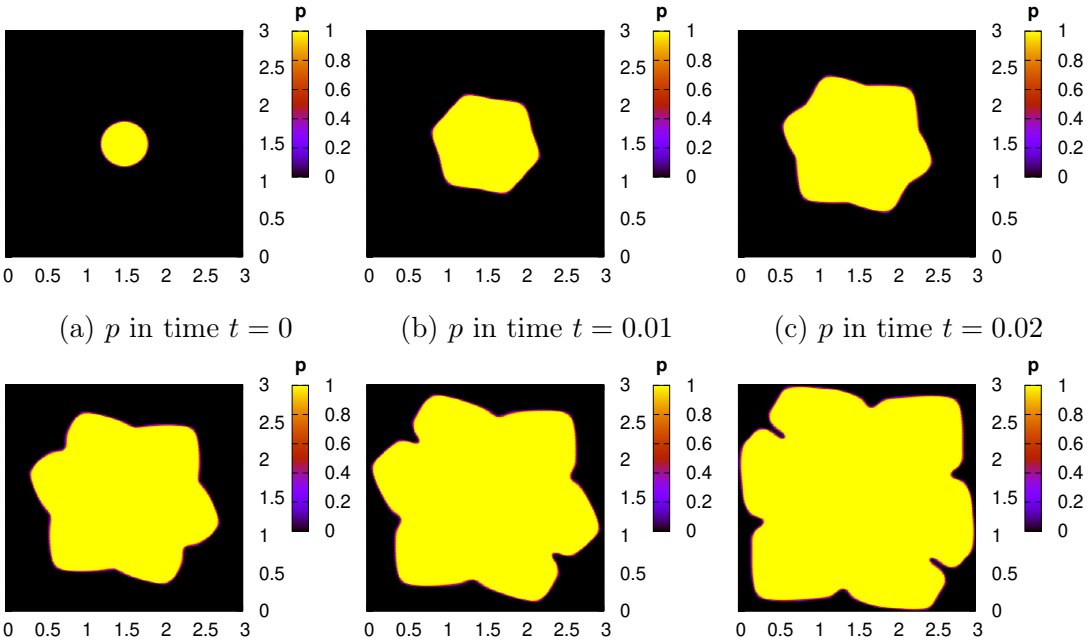
(d) p in time $t = 0.03$ (e) p in time $t = 0.04$ (f) p in time $t = 0.05$

Figure 5: The time development of a crystal growth with four dendrites, rotated by Θ_0.

5 the time development can be seen – the crystal grows starting from a round nucleus. However, unlike in the previous experiment it now grows faster in four directions. It can be noticed that these directions are rotated from diagonal of the rectangular domain by angle Θ_0. The numerical parameters are the same as in the previous experiment and can be found in Table 6.

3.2.3 Growth with Six Dendrites

This experiment is analogous to the previous one. All parameter values are identical to those in the previous experiment except those connected to anisotropy – namely A and m. m is now set to six, which further reduces the value A due to (6). This leads to a change in the number of dendrites from four to six which can be seen in Figure 6, where the time development is shown. We can notice, that in this case the solution is not symmetric. This is probably due to the insufficient precision of the numerical scheme – a usage of a more precise method will probably be required. The parameter values can



(a) p in time $t = 0$ (b) p in time $t = 0.01$ (c) p in time $t = 0.02$

(d) p in time $t = 0.03$ (e) p in time $t = 0.04$ (f) p in time $t = 0.05$

Figure 6: The time development of a crystal growth with six dendrites, rotated by Θ_0.

again be found in Table 5 in its last row. The numerical parameters are the same as in the two previous experiments and are listed in Table 6.

4 Conclusions

The main goal of this paper was to study the phase transitions in a medium. After a brief introduction, the used mathematical model was presented. This model was discretized using the finite-difference method. The numerical scheme was first used to obtain a numerical solution for problems with a known analytical solution. This helped to verify the numerical scheme. Then an anisotropy was introduced to the model and it was used to model a growth of crystals with four or six dendrites. The future work will include introduction of a more precise numerical scheme to eliminate the influence of the numerical mesh on the symmetry of the crystals and addition of obstacles into the medium to simulate a growth of crystals in a porous medium.

References

[1] M. Beneš. *Modelling of dendritic growth in pure substances.* Acta Technica CSAV Vol. 39 (1994), 375–397.

[2] M. Beneš. *Mathematical analysis of phase-field equations with numerically efficient coupling terms.* Interfaces and Free Boundaries Vol. 3 (2001), 201–221.

[3] M. Beneš. *Computational Studies of Anisotropic Diffuse Interface Model of Microstructure Formation in Solidification.* Acta Mathematica Universitatis Comenianae, Vol. 76, 1 (2007), 39–50.

[4] G. Caginalp. *An analysis of a phase field model of free boundary.* Arch. Rational Mech. Anal. Vol. 92 (1986), 92:205–245.

[5] J. Kantner. *Mathematical Model of Signal Propagation in Excitable Media.* Diploma thesis (2018), FNSPE CTU in Prague.

[6] A. Schmidt. *Computation of Three Dimensional Dendrites with Finite Elements.* Journal of Computational Physics Vol. 125, 2 (1996), 293–312.

[7] J.A. Warren, W.J. Boettinger. *Prediction of dendritic growth and microsegregation patterns in a binary alloy using the phase-field method.* Acta Metallurgica et Materialia, Vol. 43, 2 (1995), 689–703.

[8] A. Žák, M. Beneš, T. H. Illangasekare, A. C. Trautz. *Mathematical Model of Freezing in a Porous Medium at Micro-Scale.* Communications in Computational Physics Vol. 24, 2 (2019), 557–575.

[9] A. Žák, M. Beneš. *Micro-Scale Model of Thermomechanics in Solidifying Saturated Porous Media.* Acta Physics Polonica A, Vol. 134, 3 (2018), 678–682.

Performance Bound for Blind Extraction of Non-Gaussian Complex-Valued Vector Component from Gaussian Background*

Václav Kautský

4th 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. Independent Vector Extraction aims at the joint blind source extraction of K dependent signals of interest (SOI) from K linear mixtures (one signal from one mixture). Within each mixture, the signal of interest is assumed to be independent of the other signals, so SOIs form so-called independent vector component.

Compared to Independent Vector Analysis (IVA), the (de-)mixing model contains a minimum number of parameters needed for extraction, since it aims only on extraction of one vector component. The SOIs are assumed to be non-Gaussian or noncircular Gaussian, while the other signals, referred to as background, are modeled as circular Gaussian.

A Cramér–Rao–Induced Bound (CRIB) for Interference–to–Signal Ratio (ISR) is derived and compared with the similar bound for Independent Component Extraction (ICE) and IVA. The derived bound is proved to be lower or equal than that one for ICE. Numerical simulations show a good correspondence between the empirical results and the theory. The advantage based on dependency of mixtures results to better separation accuracy and helps to solve the permutation ambiguity.

Keywords: Blind Source Separation, Cramér-Rao Lower Bound, Independent Component Analysis, Independent Vector Analysis

Abstrakt. Cílem metody nazvané Independent Vector Extraction je zároveň separovat K statisticky závislých signálů (tzv. SOI) z K lineárních směsí (jeden signál z každé směsi). V dané směsi se předpokládá, že SOI je nezávislý s ostatními signály. Všechny SOI z jednotlivých směsí společně tvoří nezávislou vektorovou komponentu.

V porovnání s Independent Vector Analysis (IVA) obsahuje (de-)mixující model pouze minimální potřebný počet parametrů pro extrakci, jelikož cílem je extrakce pouze jedné komponenty. Předpokladem je, SOI signály jsou negaussovské nebo necirkulární Gaussovské.

Odvozená Crámerova-Raova dolní mez je následně porovnána s již známými mezemi pro Independent Component Extraction (ICE) a IVA. Podařilo se také ukázat, že tato mez je vždy menší než odpovídající mez pro ICE. Numerické simulace potvrzují odvozené teoretické výsledky. Využití vzájemné závislosti mezi cílovými signály SOI umožňuje dosažení vyšší separační přesnosti a také pomáhá k odstranění permutačního problému.

*This work was supported by The Czech Science Foundation through Project No. 17-00902S, and by the United States Department of the Navy, Office of Naval Research Global, through Project No. N62909-18-1-2040, and by the grant SGS18/188/OHK4/3T/14.

Klíčová slova: Analýza nezávislých komponent, Analýza nezávislých vektorů, Cramérova-Raova dolní mez, Slepá separace signálu

Full paper: V. Kautský, Z. Koldovský, and P. Tichavský. *Performance Bound for Blind Extraction of Non-Gaussian Complex-valued Vector Component from Gaussian Background.* ICASSP 2019, 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, 5287–5291.

Iterative Wiener Filtering for Deconvolution with Ringing Artifact Suppression*

Tomáš Kerepecký

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

advisor: Filip Šroubek, Department of Image Processing
Institute of Information Theory and Automation, CAS

Abstract. Sensor and lens blur degrade images acquired by digital cameras. Simple and fast removal of blur using linear filtering, such as Wiener filter, produces results that are not acceptable in most of the cases due to ringing artifacts close to image borders and around edges in the image. More elaborate deconvolution methods with non-smooth regularization, such as total variation, provide superior performance with less artifacts, however at a price of increased computational cost. We consider the alternating directions method of multipliers, which is a popular choice to solve such non-smooth convex problems, and show that individual steps of the method can be decomposed to simple filtering and element-wise operations. Filtering is performed with two sets of filters, called restoration and update filters, which are learned for the given type of blur and noise level with two different learning methods. The proposed deconvolution algorithm is implemented in the spatial domain and can be easily extended to include other restoration tasks such as demosaicing and super-resolution. Experiments demonstrate performance of the algorithm with respect to the size of learned filters, number of iterations, noise level and type of blur.

Keywords: Wiener filter, LMMSE, deconvolution, total variation, ADMM, non-smooth optimization

Abstrakt. Vlivem nedokonalého senzoru a objektivu dochází k poškození snímků pořízených digitálními fotoaparáty. K odstranění těchto rozmazání je možné použít lineární filtry, např. Wienerův filtr, nicméně výsledky jsou obvykle neuspokojivé vzhledem k výskytu tzv. Gibsových artefaktů v blízkosti hran a okrajů obrázku. Lepších výsledků s menším výskytem artefaktů je možné dosáhnout při použití propracovaněších dekonvolučních metod s využitím nehladké regularizace, např. totální variace, ovšem za cenu vyšší výpočetní náročnosti. V tomto článku uvažujeme numerickou metodu ADMM (Alternating Directions Method of Multipliers), řešící nehladké konvexní problémy a ukazujeme, že jednotlivé kroky této metody je možné rozložit na jednoduché filtrování a prahování po prvcích. Filtrování je provedeno za pomocí dvou skupin filtrů: rekonstrukční a aktualizační, které jsou naučeny pro daný typ rozmazání a úrovně šumu za pomocí dvou různých učících mechanizmů. Navrhovaný dekonvoluční algoritmus je implementován v obrazové doméně a je možné ho snadno rozšířit o další rekonstrukční problémy, jako demosaicing či super-resolution. Provedené experimenty demonstrují kvalitu rekonstrukce v závislosti na velikosti naučených filtrů, počtu iterací, úrovni šumu a typu rozmazání.

*This work has been supported by the grant: *Řešení inverzních problémů vznikajících při analýze rychle se pohybujících objektů* GAČR: GA18-05360S.

Klíčová slova: Wienerův filtr, LMMSE, dekonvoluce, totalní variace, ADMM, nehladká optimalizace

Full paper: Filip Šroubek, Tomáš Kerepecký and Jan Kamenický. *Iterative Wiener Filtering for Deconvolution with Ringing Artifact Suppression.* In: 27th European Signal Processing Conference. A Coruña, Spain, September 2–6, 2019.

References

[1] M. Almeida and M. Figueiredo. *Deconvolving images with unknown boundaries using the alternating direction method of multipliers.* Image Processing, IEEE Transactions on **22** (Aug 2013), 3074–3086.

[2] M. R. Banham and A. K. Katsaggelos. *Digital image restoration.* **14** (March 1997), 24–41.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. *Distributed optimization and statistical learning via the alternating direction method of multipliers.* Foundations and Trends® in Machine learning **3** (2011), 1–122.

[4] A. Chambolle and T. Pock. *A first-order primal-dual algorithm for convex problems with applications to imaging.* Journal of mathematical imaging and vision **40** (2011), 120–145.

[5] A. Chambolle and T. Pock. *An introduction to continuous optimization for imaging.* Acta Numerica **25** (2016), 161–319.

[6] R. Liu and J. Jia. Reducing boundary artifacts in image deconvolution. In 'Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on', 505–508. IEEE, (2008).

[7] J. Portilla. Maximum likelihood extension for non-circulant deconvolution. In 'Proc. IEEE Int. Conf. Image Processing (ICIP)', 4276–4279, (October 2014).

[8] J. Portilla, D. Otaduy, and C. Dorronsoro. Low-complexity linear demosaicing using joint spatial-chromatic image statistics. In 'Proc. IEEE Int. Conf. Image Processing 2005', volume 1, I–61, (September 2005).

[9] S. J. Reeves. *Fast image restoration without boundary artifacts.* **14** (2005), 1448–1453.

[10] L. Rudin, S. Osher, and E. Fatemi. *Nonlinear total variation based noise removal algorithms.* Physica D **60** (1992), 259–268.

[11] M. Šorel. *Removing boundary artifacts for real-time iterated shrinkage deconvolution.* IEEE Transactions on Image Processing **21** (Apr 2012), 2329–2334.

Parallel Implementation of Immersed Boundary–Lattice Boltzmann Method 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. The lattice Boltzmann method (LBM) is an efficient numerical method capable of simulating fluid flow in the laminar as well as turbulent regime. LBM is suitable for massive parallelization on GPUs, but domain decomposition and using multiple GPUs is needed for high-resolution simulations due to limited amount of memory on a single GPU. In order to simulate flow over complex immersed bodies or interactions with elastic structures, it can be coupled with the immersed boundary method (IBM). In our recent work, we proposed a modification of the method to improve the conditioning of the linear systems, which leads to faster convergence. In this work, we focus on the implementation of LBM for GPU clusters and investigate the efficient implementation of IB-LBM on a single GPU only. We consider several numerical methods, preconditioners, and libraries such as CUSPARSE and TNL and present the impact of each approach on the overall performance of the IB-LBM solver.

Keywords: lattice Boltzmann method, immersed boundary method, parallel implementation on GPU, computational study

Abstrakt. Mřížková Boltzmannova metoda (LBM) je efektivní numerická metoda schopná simulovat proudění tekutin v laminárním i turbulentním režimu. LBM je vhodná pro masivní paralelizaci na GPU akcelerátorech, ale pro simulace ve vysokém rozlišení je kvůli omezenému množství paměti na GPU potřeba využít více akcelerátorů. Pro simulaci proudění kolem těles s komplikovanou hranicí nebo interakce s elastickými tělesy je možné zkombinovat LBM a metodu vnořené hranice (IBM). V naší nedávné práci jsme navrhli modifikaci této metody zlepšující podmíněnost soustavy lineárních rovnic, což vede k rychlejší konvergenci. V této práci se zaměřujeme na implementaci LBM pro klastry využívající GPU akcelerátory a zkoumáme efektivitu implementace metody IB-LBM na jednom GPU. Uvažujeme několik numerických metod, předpodmínění a knihoven jako CUSPARSE a TNL a porovnáváme vliv těchto přístupů na celkový výkon řešiče IB-LBM.

Klíčová slova: mřížková Boltzmannova metoda, metoda vnořené hranice, paralelní implementace na GPU, výpočetní studie

1 Lattice Boltzmann method

The lattice Boltzmann method (LBM) is a popular computational method for the simulation of fluids. Instead of solving the Navier–Stokes equations, LBM is based on approximating the temporal evolution of macroscopic quantities such as density ρ, velocity $\boldsymbol{v}$, and others (pressure, stress tensor, etc.) using probability moments of discrete particle distribution functions $f_k = f_k(\boldsymbol{x}, t)$, $k = 1, \dots, q$, where q denotes the number of discrete velocities per lattice site. In this work, we consider a regular 3D lattice with the D3Q27 model consisting of $q = 27$ discrete velocities per lattice site denoted as $\boldsymbol{\xi}_k$, $k = 1, \dots, q$. The evolution of f_k is described by the following discrete Boltzmann transport equation for all $k = 1, \dots, q$,

$$f_k(\boldsymbol{x}_\iota + \Delta t \boldsymbol{\xi}_k, t + \Delta t) - f_k(\boldsymbol{x}_\iota, t) = \mathcal{C}_k(\boldsymbol{x}_\iota, t) + \mathcal{G}_k(\boldsymbol{x}_\iota, t), \quad \forall \boldsymbol{x}_\iota \in \hat{\Omega},\ \forall t > 0, \tag{1}$$

where $\hat{\Omega}$ denotes the set of all discrete lattice sites, $\mathcal{C}_k$ denotes the discrete collision operator, $\mathcal{G}_k$ denotes the discrete forcing term, and Δt denotes the dimensionless time interval of one LBM iteration, which is set to unity, i.e., $\Delta t = 1$. Macroscopic quantities can be reconstructed by taking moments of the discrete particle distribution functions f_k, for example the macroscopic density ρ is given by

$$\rho(\boldsymbol{x}_\iota, t) = \sum_{k=1}^{q} f_k(\boldsymbol{x}_\iota, t) \tag{2}$$

and the macroscopic velocity $\boldsymbol{v}$ can be computed from the macroscopic momentum density $\rho\boldsymbol{v}$ which is given by

$$\rho(\boldsymbol{x}_\iota, t)\boldsymbol{v}(\boldsymbol{x}_\iota, t) = \sum_{k=1}^{q} f_k(\boldsymbol{x}_\iota, t)\boldsymbol{\xi}_k + \frac{1}{2}\boldsymbol{g}(\boldsymbol{x}_\iota, t)\Delta t, \tag{3}$$

where $\boldsymbol{g}$ denotes the external force density exerted on the fluid. The detailed derivation of LBM is out of scope of this work and we refer the reader to [7] for details.

1.1 Computational algorithm

The main advantage of LBM is that the computational algorithm for solving Eq. (1) can be split into a collision and a streaming step. In the collision step, $\mathcal{C}_k$ and $\mathcal{G}_k$ are evaluated using only local values of f_k and in the streaming step the value of each f_k is transferred in the direction of $\boldsymbol{\xi}_k$ to the neighbouring lattice site for $k = 1, \dots, q$. The LBM algorithm can be formulated in several ways with respect to the streaming step, with the so-called *push* and *pull schemes* being the most straightforward ones. In the push scheme, the streaming step follows the collision step and the post-collision values of each f_k are written (i.e., pushed) to the corresponding neighbouring lattice sites. On the other hand, the pull scheme orders streaming before the collision step and thus the pre-collision values of each f_k are read (i.e., pulled) from the relevant neighbouring lattice sites. The algorithm using the pull scheme is summarized below in Algorithm 1.

Algorithm 1 (LBM)

1. Initialization (read input data, set initial condition, etc.)
2. While the final time is not reached, do for all $\boldsymbol{x}_\iota \in \hat{\Omega}$:

2.1. Streaming step (read the pre-collision state of all f_k from neighbouring sites).
2.2. If $\boldsymbol{x}_\iota$ lies on the domain boundary, handle boundary conditions.
2.3. Collision step (evaluate the collision operator $\mathcal{C}_k$ and forcing term $\mathcal{G}_k$).
2.4. Output data (evaluate and write all predetermined macroscopic quantities).

A disadvantage of the push and pulled schemes is that in order to avoid write conflicts in the implementation of the LBM algorithm, two sets of distribution functions f_k are needed – one for the current time level and another for the next time level. On the other hand, this so-called *A-B pattern* can be easily implemented as well as parallelized. Alternative approaches, such as the *A-A pattern* [1, 12] or Esoteric Twist algorithm [6], perform the streaming step in-place and thus need only one array for storing the values of f_k at a single time level, but their implementation is more complicated. In this work, we consider only the pull scheme and the A-B pattern.

In order to utilize multiple GPUs or even multiple nodes for computation, the LBM algorithm has to be reformulated following the domain decomposition approach. In short, the computational domain Ω is split into subdomains Ω_j, $j = 1, \ldots, N_p$ which are processed independently using N_p processors and after each iteration, the necessary data are copied between neighbouring subdomains in order to provide data computed elsewhere. This approach is summarized below in Algorithm 2. However, a naïve implementation of this domain decomposition algorithm does not scale with increasing number of GPUs, which can be seen from the results of weak scaling analysis in Table 1. The benchmark problem was computed on the RCI cluster which contains 4 Nvidia Tesla V100 GPUs per node (i.e., 48 GPUs in total).

Algorithm 2 (Distributed LBM)

1. Initialization (read input data, set initial condition, etc.)
2. Copy distribution functions on the boundaries between subdomains.
3. While the final time is not reached:

3.1. For all lattice sites on all subdomains, perform steps 2.1-2.4 of Algorithm 1.
3.2. Copy distribution functions on the boundaries between subdomains.

N_{nodes}	N_{GPUs}	GLUPS	*Eff*
1	1	2.5	1.00
1	4	7.8	0.79
2	8	11.5	0.58
4	16	14.0	0.35

Table 1: Performance of the naïve implementation of distributed LBM (Algorithm 2).

In order to improve the scaling of the distributed LBM algorithm, several optimizations are necessary. First of all, it is necessary to *overlap* computation with communication in order to hide the extra latency which was not present in the original Algorithm 1. This can be done by processing the boundary lattice sites of each subdomain separately from the interior lattice sites, copying the data between subdomains as soon as the computation on the boundary has finished, and processing the interior lattice sites independently while all processes exchange the boundary data. This approach can be summarized in Algorithm 3 using *asynchronous* operations for communication between processes.

Algorithm 3 (Distributed LBM with overlapped computation and communication)

1. Initialization (read input data, set initial condition, etc.)
2. Copy distribution functions on the boundaries between subdomains.
3. While the final time is not reached:

3.1. On all subdomains, start processing *boundary lattice sites*.
3.2. On all subdomains, start processing *interior lattice sites*.
3.3. On all subdomains, wait until boundary lattice sites are processed.
3.4. On all subdomains, initiate *asynchronous communication* for boundaries.
3.5. Wait for all preceding operations to finish.

Another important optimization is that the boundary data should be copied without using custom buffers. This is possible in general only with a 1D distribution of the processes (i.e., processes are numbered linearly and each has at most two neighbours: one on the left and one on the right) and with an appropriate storage layout for the underlying multidimensional array in order to ensure contiguity of the transferred data. Finally, for the LBM algorithm it is not necessary to exchange all 27 distribution functions f_k between each neighbouring processes, but the communication size can be reduced to only 9 specific distribution functions which should be streamed from one subdomain to another. However, a specific ordering of the distribution functions f_k is necessary due to the previous optimization.

These optimizations can be implemented with a so-called *CUDA-aware* MPI library and using CUDA streams for the management of asynchronous operations on the GPU. The resulting performance of the optimized LBM solver on the RCI cluster is shown in Table 2. For 8 nodes (32 GPUs) there is still a significant drop in the parallel efficiency, but this is likely caused by a hardware problem (the cluster has a star topology for the high-speed InfiniBand network and the central switch may be saturated).

2 Immersed bounary–lattice Boltzmann method

In this section, we describe the modified implicit immersed boundary–lattice Boltzmann method (IB-LBM) based on our recent paper [5] and present the implementation details as well as results of a computational performance study performed on a modern GPU system.

The immersed boundary method (IBM) introduced by C. Peskin in [11] allows simulating the fluid–structure interaction using Eulerian coordinates for the fluid description

N_{nodes}	N_{GPUs}	GLUPS	*Eff*
1	1	2.5	1.00
1	4	10.4	1.05
2	8	19.3	0.97
4	16	39.4	0.99
8	32	50.1	0.63

Table 2: Performance of the optimized implementation of distributed LBM with overlapped computation and communication, unbuffered communication and reduced communication size.

and Lagrangian coordinates for the description of the immersed body (elastic) boundary. Similarly to [15, 4, 10, 3], our overall motivation for using the IBM is its coupling with the Lattice Boltzmann Method (LBM) and developing an efficient numerical solver on GPU which, for instance, can be used to investigate blood flow patterns in large arteries with non-rigid walls. In most cases, however, the coupled immersed boundary–lattice Boltzmann method (IB-LBM) does not guarantee impermeability of the discretized body boundary. Therefore, in order to prevent penetrative fluid flow through the boundary, the choice of the Lagrangian discretization requires careful treatment and further investigation. In our recent paper [5], we investigated the effects of the spacing of Lagrangian nodes discretizing a rigid and immobile immersed body boundary and based on [13, 14], we proposed a modified implicit IB-LBM which significantly improves the conditioning of linear systems for densely-spaced discretizations.

From the performance point of view, solving large linear systems of equations resulting from the implicit IB-LBM is the most crucial step. In this section, we compare several numerical methods, preconditioners, and libraries such as CUSPARSE [9] and TNL, and investigate the impact of each approach on the overall performance.

2.1 Mathematical and numerical models

Solely for the purpose of the computational study presented here, we consider flow in three-dimensional domains without gravity, the fluid is assumed as Newtonian and incompressible, and the immersed body is considered rigid and immobile. The mathematical model describing the fluid flow and the immersed body–fluid interaction is represented by the following set of partial differential equations [11]:

$$\rho(\boldsymbol{x},t)\left(\frac{\partial \boldsymbol{v}(\boldsymbol{x},t)}{\partial t}+\boldsymbol{v}(\boldsymbol{x},t)\cdot\nabla\boldsymbol{v}(\boldsymbol{x},t)\right)+\nabla p(\boldsymbol{x},t)=\mu(\boldsymbol{x},t)\Delta\boldsymbol{v}(\boldsymbol{x},t)+\boldsymbol{g}(\boldsymbol{x},t), \tag{4a}$$

$$\nabla\cdot\boldsymbol{v}(\boldsymbol{x},t)=0, \tag{4b}$$

$$\boldsymbol{g}(\boldsymbol{x},t)=\int_{\Gamma}\boldsymbol{g}_b\left(\boldsymbol{X}(\boldsymbol{s},t),t\right)\delta\left(\boldsymbol{x}-\boldsymbol{X}(\boldsymbol{s},t)\right)\mathrm{d}\boldsymbol{s}, \tag{4c}$$

$$\frac{\partial \boldsymbol{X}}{\partial t}(\boldsymbol{s},t)=\boldsymbol{v}_b\left(\boldsymbol{X}(\boldsymbol{s},t),t\right)=\int_{\Omega}\boldsymbol{v}(\boldsymbol{x},t)\delta\left(\boldsymbol{x}-\boldsymbol{X}(\boldsymbol{s},t)\right)\mathrm{d}\boldsymbol{x}, \tag{4d}$$

for all time $t > 0$, all Eulerian spatial coordinates $\boldsymbol{x} \in \Omega$, where $\Omega \subset \mathbb{R}^3$ denotes the computational domain, and all $\boldsymbol{s}$ that denote the parametric coordinate of the Lagrangian points $\boldsymbol{X} \in \Gamma$, where Γ is the boundary of the immersed body. In Eq. (4), ρ is the fluid density, $\boldsymbol{v}$ and $\boldsymbol{v}_b$ are the fluid velocities in the Eulerian and Lagrangian description, respectively, p is the fluid pressure, μ is the dynamic viscosity of the fluid, $\boldsymbol{g}$ and $\boldsymbol{g}_b$ are the force density source terms realizing the fluid-body interaction in the Eulerian and Lagrangian description, respectively, and δ denotes the Dirac delta function. Since δ is not a regular function, Eq. (4) holds in the weak sense for generalized functions.

The numerical scheme uses LBM to solve the Navier–Stokes equations given by Eqs. (4a) and (4b) with the forcing term $\boldsymbol{g}$ given by Eq. (4c), see Section 1 for details. In general, the temporal evolution of the body boundary in the Lagrangian coordinates given by Eq. (4d) can be treated, e.g., with the finite difference method. However, in the special case of an immobile body boundary, we have $\frac{\partial \boldsymbol{X}}{\partial t}(\boldsymbol{s}, t) = 0$ and the Lagrangian coordinates stay constant. The body boundary Γ is discretized by a finite set $\hat{\Gamma}$ consisting of N_L Lagrangian points $\boldsymbol{X}_\ell \in \hat{\Gamma}$, conveniently numbered for further derivation as $\ell = 1, \dots, N_L$. The computational domain Ω, assumed rectangular, is discretized by a finite lattice $\hat{\Omega}$ (a regular rectangular grid) represented by Eulerian points $\boldsymbol{x}_\iota \in \hat{\Omega}$, conveniently numbered as $\iota = 1, \dots, N_E$.

Based on [5], Eqs. (4c) and (4d) are discretized as follows. The fluid velocity $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t)$ at some Eulerian point $\boldsymbol{x} \in \Omega$ and time t that is assumed in the form

$$\boldsymbol{v}(\boldsymbol{x}, t) = \boldsymbol{v}^*(\boldsymbol{x}, t) + \boldsymbol{c}(\boldsymbol{x}, t), \tag{5}$$

where $\boldsymbol{c}$ is the unknown velocity correction and $\boldsymbol{v}^*$ denotes the intermediate velocity that is in the discrete form given by

$$\boldsymbol{v}^*(\boldsymbol{x}_\iota, t) = \frac{1}{\rho(\boldsymbol{x}_\iota, t)} \sum_{k=1}^{q} f_k(\boldsymbol{x}_\iota, t) \boldsymbol{\xi}_k, \tag{6}$$

for all $\boldsymbol{x}_\iota \in \hat{\Omega}$. Once the velocity correction $\boldsymbol{c}(\boldsymbol{x}_\iota, t)$ in the discrete form is known, the desired force density $\boldsymbol{g}$ is given by

$$\boldsymbol{g}(\boldsymbol{x}_\iota, t) = \frac{2\rho(\boldsymbol{x}_\iota, t)}{\Delta t} \boldsymbol{c}(\boldsymbol{x}_\iota, t). \tag{7}$$

The equation for $\boldsymbol{c}$ is derived in the continuous spaces Ω and Γ using the properties of convolutions with the Dirac delta function. The body boundary velocity $\boldsymbol{v}_b$ at some Lagrangian point $\boldsymbol{X}(\boldsymbol{r}) \in \Gamma$ is related to the decomposed velocity $\boldsymbol{v}$ from Eq. (5) in the Eulerian description as

$$\boldsymbol{v}_b(\boldsymbol{X}(\boldsymbol{r}), t) = \int_\Omega \Big(\boldsymbol{v}^*(\boldsymbol{x}, t) + \boldsymbol{c}(\boldsymbol{x}, t) \Big) \delta(\boldsymbol{x} - \boldsymbol{X}(\boldsymbol{r})) \mathrm{d}\boldsymbol{x}. \tag{8}$$

Analogously, $\boldsymbol{c}$ is related to the unknown boundary velocity correction $\boldsymbol{c}_b$ in the Lagrangian description as

$$\boldsymbol{c}(\boldsymbol{x}, t) = \int_\Gamma \boldsymbol{c}_b(\boldsymbol{X}(\boldsymbol{s}), t) \delta(\boldsymbol{x} - \boldsymbol{X}(s)) \mathrm{d}\boldsymbol{s}. \tag{9}$$

Combining Eqs. (8) and (9) and using the properties of the Dirac delta function, we obtain

$$\boldsymbol{v}_b(\boldsymbol{X}(\boldsymbol{r},t),t) = \int_{\Omega} \boldsymbol{v}^*(\boldsymbol{x},t)\delta(\boldsymbol{x}-\boldsymbol{X}(\boldsymbol{r}))\mathrm{d}\boldsymbol{x} + \int_{\Gamma} \boldsymbol{c}_b(\boldsymbol{X}(\boldsymbol{s}),t)\delta(\boldsymbol{X}(\boldsymbol{r})-\boldsymbol{X}(\boldsymbol{s}))\mathrm{d}\boldsymbol{s}. \quad (10)$$

The first integral in Eq. (10) mediates the transfer between the Lagrangian and Eulerian coordinates and the second integral is effective only in the Lagrangian description. Equation (10) can be discretized as

$$\boldsymbol{v}_b(\boldsymbol{X}_\ell,t) = \sum_{\iota=1}^{N_E} \boldsymbol{v}^*(\boldsymbol{x}_\iota,t)D(\boldsymbol{x}_\iota-\boldsymbol{X}_\ell)\Delta\boldsymbol{x}_\iota + \sum_{k=1}^{N_L} \boldsymbol{c}_b(\boldsymbol{X}_k,t)D(\boldsymbol{X}_\ell-\boldsymbol{X}_k)\Delta\boldsymbol{s}_k, \quad (11)$$

for all $\boldsymbol{X}_\ell \in \hat{\Gamma}$, where $\Delta\boldsymbol{x}_\iota = h^3$ is the volume of the dual volume centred around $\boldsymbol{x}_\iota \in \hat{\Omega}$, $\Delta\boldsymbol{s}_\ell \in \mathbb{R}$ denotes the size of a boundary element around $\boldsymbol{X}_\ell$, and D denotes the continuous kernel distribution proposed by C. Peskin in [11] defined by

$$D(\boldsymbol{z}) = \prod_{k=1}^{3} \frac{1}{h}\phi\left(\frac{z^{(k)}}{h}\right), \quad (12)$$

where $z^{(k)}$ denotes the k-th component of a vector $\boldsymbol{z}$ and ϕ is a continuous approximation of the one-dimensional Dirac delta function [8]. According to [5], a simple choice of ϕ given by

$$\phi(r) = \begin{cases} 1-|r| & \text{for } |r| \le 1, \\ 0 & \text{for } |r| \ge 1, \end{cases} \quad (13)$$

leads to reasonable results for all choices of the Lagrangian discretization.

The coefficients $\mathbf{A}_{\ell,k} = D(\boldsymbol{X}_\ell - \boldsymbol{X}_k)$ form a sparse, symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{N_L,N_L}$. Hence, Eq. (11) represents a linear system of equations for $\boldsymbol{c}_b(\boldsymbol{X}_k,t)\Delta\boldsymbol{s}_k$, i.e., for the components of the unknown boundary velocity correction $\boldsymbol{c}_b(\boldsymbol{X}_k,t)$ multiplied by the size of the boundary element $\Delta\boldsymbol{s}_k$, $k = 1,2,\ldots,N_L$. Once $\boldsymbol{c}_b(\boldsymbol{X}_k,t)\Delta\boldsymbol{s}_k$ is known, the velocity correction $\boldsymbol{c}(\boldsymbol{x}_\iota,t)$ is obtained for all $\boldsymbol{x}_\iota \in \hat{\Omega}$ from the discrete form of Eq. (9) as

$$\boldsymbol{c}(\boldsymbol{x}_\iota,t) = \sum_{k=1}^{N_L} D(\boldsymbol{x}_\iota - \boldsymbol{X}_k)\boldsymbol{c}_b(\boldsymbol{X}_k,t)\Delta\boldsymbol{s}_k, \quad (14)$$

i.e., the computation of $\Delta\boldsymbol{s}_k$ is actually not necessary in order to obtain the velocity correction $\boldsymbol{c}(\boldsymbol{x}_\iota,t)$ in the Eulerian description.

2.2 Computational algorithm

Based on the preceding derivation of the numerical method, the computational algorithm can be summarized in Algorithm 4.

Algorithm 4 (IB-LBM)
For each time level t_j:

1. Compute $\boldsymbol{v}^*(\boldsymbol{x}_\iota, t_j)$ for all lattice sites $\boldsymbol{x}_\iota \in \hat{\Omega}$ using Eq. (6).
2. Compute the first sum in Eq. (11) for all $\boldsymbol{X}_\ell \in \hat{\Gamma}$, i.e., $\boldsymbol{v}_b^*(\boldsymbol{X}_\ell, t_j) = \sum_{\iota=1}^{N_E} \mathbf{B}_{\ell,\iota} \boldsymbol{v}^*(\boldsymbol{x}_\iota, t_j) \Delta \boldsymbol{x}_\iota$, where $\mathbf{B}_{\ell,\iota} = D(\boldsymbol{x}_\iota - \boldsymbol{X}_\ell)$.
3. For $i = 1, 2, 3$, denote $\boldsymbol{b}_i \equiv [-\boldsymbol{v}_b^{*,(i)}(\boldsymbol{X}_\ell, t_j)]_{\ell=1}^{N_L} \in \mathbb{R}^{N_L}$ and $\boldsymbol{y}_i \equiv [\boldsymbol{c}_b^{(i)}(\boldsymbol{X}_\ell, t_j) \Delta \boldsymbol{s}_\ell]_{\ell=1}^{N_L} \in \mathbb{R}^{N_L}$, where the upper index $^{(i)}$ denotes the i-th component of the preceding vector, and solve the linear system $\mathbf{A}\boldsymbol{y}_i = \boldsymbol{b}_i$ (c.f. Eq. (11) where we set $\boldsymbol{v}_b(\boldsymbol{X}_\ell, t) \equiv 0$).
4. Compute the velocity correction $\boldsymbol{c}(\boldsymbol{x}_\iota, t_j)$ for all $\boldsymbol{x}_\iota \in \hat{\Omega}$ using Eq. (14), i.e., $\boldsymbol{c}(\boldsymbol{x}_\iota, t_j) = \sum_{k=1}^{N_L} \mathbf{B}_{k,\iota} \boldsymbol{c}_b(\boldsymbol{X}_k, t) \Delta \boldsymbol{s}_k$.
5. Compute the force density $\boldsymbol{g}(\boldsymbol{x}_\iota, t_j)$ for all $\boldsymbol{x}_\iota \in \hat{\Omega}$ using Eq. (7).
6. Perform an LBM iteration (i.e., steps 2.1-2.4 of Algorithm 1).

Note that in the Step 2, the multiplication with a matrix $\mathbf{B} = [\mathbf{B}_{\ell,\iota}]_{\ell=1,\iota=1}^{N_L,N_E}$ mediates the transformation from Eulerian to Lagrangian coordinates, and in the Step 4, the multiplication with a transposed matrix $\mathbf{B}^T$ mediates the transformation from Lagrangian to Eulerian coordinates.

2.3 Implementation strategies

The IB-LBM computational algorithm contains several steps for which the most suitable implementation strategy is not obvious. Firstly, as described in Section 1, the LBM algorithm alone is suitable for the implementation on GPUs, so we do not consider a CPU-only implementation of IB-LBM. On the other hand, the solution of the linear systems in the Step 3 of Algorithm 4 is easily done on CPU using an external library such as UMFPACK [2]. However, combining the computation of LBM on the GPU and the solution of linear systems on the CPU would cause a lot of communication due to data transfers between the GPU and CPU in every time step, which might lead to bad performance. Hence, additional approaches reducing the amount of data transfers or moving the solution of the linear systems to the GPU need to be examined.

For the computational study presented below, we consider the following approaches. The linear systems from the Step 3 of Algorithm 4 are solved either with UMFPACK [2] on CPU or with the conjugate gradients (CG) method either on CPU or GPU. We compare two implementations of the CG method: TNL provides an implementation for both CPU and GPU and CUSPARSE [9] provides an implementation for GPU only. The TNL-based implementation of the CG method is used either without a preconditioner or with the incomplete Cholesky factorization (IC(0)). Finally, the coordinate transformations in Steps 2 and 4 are computed either on CPU or GPU.

2.4 Numerical results

Results of a detailed qualitative study involving the presented method have been published in [5]. In this section, we extend those results with an analysis of computational performance. The simulations presented here were computed on one node of the RCI

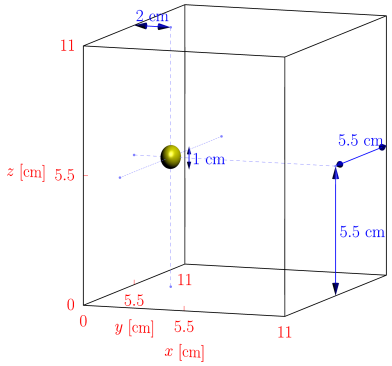


Figure 1: Setup of the computational domain for the benchmark of flow around a sphere. Inflow boundary is on the left and outflow boundary is on the right.

Implementation	MLUPS
UMFPACK	
– coordinate transformations on CPU	99
– coordinate transformations on GPU	728
Conjugate gradients (no preconditioner)	
– CUSPARSE-based (GPU)	1013
– TNL-based (GPU)	1415
– TNL-based (CPU)	1406
Conjugate gradients (IC(0) preconditioner)	
– TNL-based (GPU)	1435
– TNL-based (CPU)	1424

Table 3: Performance comparison for the flow around a sphere in the highest resolution.

cluster, which contains 2 CPUs Intel Xeon Gold 6130 (16 cores, 384 GiB RAM) and 4 GPUs NVIDIA Tesla V100 (5120 cores, 32 GiB memory). However, only 1 CPU and at most 1 GPU were used in every simulation.

As the first benchmark, we computed the flow around a rigid sphere in the computational domain depicted in Fig. 1. The inflow velocity is prescribed such that the Reynolds number based on the sphere diameter is $Re = 100$. The sphere was discretized such that the maximal distance σ between two neighbouring Lagrangian points is equal to the Eulerian lattice spacing h. In the highest resolution considered, i.e. $h = \sigma = 11/384$ cm, there are approximately 56×10^6 Eulerian points and 7620 Lagrangian points. The results in Table 3 show significant difference between coordinate transformations computed on CPU and GPU. Therefore, different strategies for the computation of coordinate transformations were considered only with the UMFPACK solver and in all following simulations, coordinate transformations were computed on the GPU. The results also show a significant advantage of the TNL-based implementation of the CG method compared to the CUSPARSE-based implementation. On the other hand, the use of an IC(0) preconditioner does not make a significant difference in the performance.

The second benchmark is the flow around a rigid cylinder in a computational domain whose 2D cross-section is depicted in Fig. 2. The inflow velocity is prescribed such that the

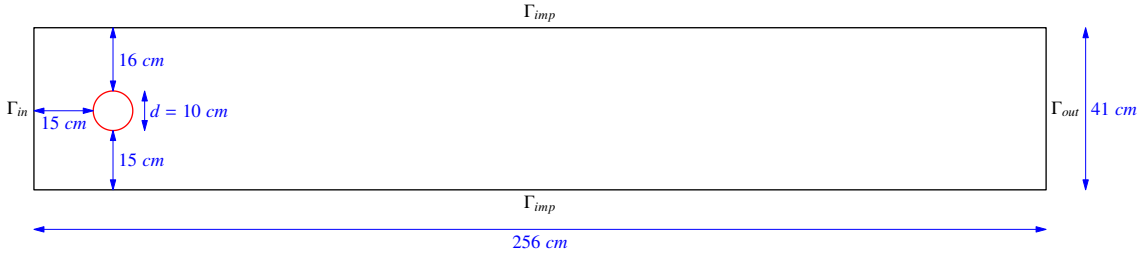


Figure 2: 2D cross-section of the computational domain for the benchmark of flow around a cylinder. Inflow boundary is on the left and outflow boundary is on the right.

Implementation	MLUPS
UMFPACK	
– coordinate transformations on CPU	85
– coordinate transformations on GPU	942
Conjugate gradients (no preconditioner)	
– CUSPARSE-based (GPU)	930
– TNL-based (GPU)	1236
– TNL-based (CPU)	904
Conjugate gradients (IC(0) preconditioner)	
– TNL-based (GPU)	1218
– TNL-based (CPU)	899

Table 4: Performance comparison for the flow around a cylinder in the highest resolution.

Reynolds number based on the cylinder diameter is $Re = 100$. Compared to a sphere used in the previous benchmark, the discretization of a cylinder spanning the whole depth of the computational domain leads to a higher ratio of the number of Lagrangian points and the number of Eulerian points. In the highest resolution considered, i.e. $h = \sigma = 41/192$ cm, there are approximately 34×10^6 Eulerian points and 55×10^3 Lagrangian points. This leads to a lower performance measured in MLUPS as shown in Table 4, but it also highlights the benefits of executing the CG method on the GPU rather than on the CPU.

3 Conclusion

We have described parallel implementation of LBM and IB-LBM for GPUs and shown results of computational studies to compare the effects of various optimizations. Our MPI-based implementation of the LBM shows good scalability up to 4 nodes (16 GPUs). The highest performance of the IB-LBM solver has been achieved by computing the coordinate transformations on the GPU and by using the TNL-based implementation of the CG method on the GPU. For the presented benchmarks with a rigid and immobile immersed bodies, the use of a strong preconditioner such as IC(0) does not seem to be necessary, because the count of iterations is already very small even without a preconditioner. For problems with moving or elastic bodies, however, a strong preconditioner may be necessary for a good performance.

4 Acknowledgement

This work has been supported by the Student Grant Agency of the Czech Technical University in Prague project no. SGS17/194/OHK4/3T/14, the Czech Science Foundation project no. 18-09539S, and the Operational Programme Research, Development and Education project no. CZ.02.1.01/0.0/0.0/16_019/0000765.

References

[1] P. Bailey, J. Myre, S. D. Walsh, D. J. Lilja, and M. O. Saar. Accelerating lattice Boltzmann fluid flow simulations using graphics processors. In 'International Conference on Parallel Processing', 550–557. IEEE, (2009).

[2] T. A. Davis. *Umfpack version 4.4 user guide.* Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL (2005).

[3] A. De Rosis and E. Lévêque. *Central-moment lattice Boltzmann schemes with fixed and moving immersed boundaries.* Computers & Mathematics with Applications **72** (2016), 1616–1628.

[4] Z.-G. Feng and E. E. Michaelides. *The immersed boundary-lattice Boltzmann method for solving fluid–particles interaction problems.* Journal of Computational Physics **195** (2004), 602–628.

[5] R. Fučík, P. Eichler, R. Straka, P. Pauš, J. Klinkovský, and T. Oberhuber. *On optimal node spacing for immersed boundary–lattice Boltzmann method in 2D and 3D.* Computers & Mathematics with Applications **77** (2019), 1144 – 1162.

[6] M. Geier and M. Schönherr. *Esoteric twist: an efficient in-place streaming algorithmus for the lattice Boltzmann method on massively parallel hardware.* Computation **5** (2017), 19.

[7] T. Krüger, H. Kusumaatmaja, A. Kuzmin, O. Shardt, G. Silva, and E. M. Viggen. *The Lattice Boltzmann Method.* Springer, (2017).

[8] T. Krüger, F. Varnik, and D. Raabe. *Efficient and accurate simulations of deformable particles immersed in a fluid using a combined immersed boundary lattice Boltzmann finite element method.* Computers & Mathematics with Applications **61** (2011), 3485–3505.

[9] NVIDIA. CUDA toolkit documentation, version 10.1, (2019).

[10] D. Owen, C. Leonardi, and Y. Feng. *An efficient framework for fluid–structure interaction using the lattice Boltzmann method and immersed moving boundaries.* International Journal for Numerical Methods in Engineering **87** (2011), 66–95.

[11] C. S. Peskin. *The immersed boundary method.* Acta numerica **11** (2002), 479–517.

[12] M. Wittmann, T. Zeiser, G. Hager, and G. Wellein. *Comparison of different propagation steps for lattice Boltzmann methods.* Computers & Mathematics with Applications **65** (2013), 924–935.

[13] J. Wu and C. Shu. *Implicit velocity correction-based immersed boundary-lattice Boltzmann method and its applications.* Journal of Computational Physics **228** (2009), 1963–1979.

[14] J. Wu and C. Shu. *An improved immersed boundary-lattice Boltzmann method for simulating three-dimensional incompressible flows.* Journal of Computational Physics **229** (2010), 5022–5042.

[15] T.-H. Wu, M. Khani, L. Sawalha, J. Springstead, J. Kapenga, and D. Qi. *A CUDA-based implementation of a fluid-solid interaction solver: the immersed boundary lattice-Boltzmann lattice-spring method.* Comput. Phys **23** (2018), 980.

Rich-Club Property of (Partial) Correlation Matrix*

Jakub Kořenek

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

advisor: Jaroslav Hlinka, Department of Complex Systems
Institute of Computer Science, CAS

Abstract. In the description of complex dynamical systems using graph theory, Pearson linear correlation is the most commonly used approach to graph construction. However, it was shown that using correlation can strongly affect the resulting network and its graph-theoretical properties. This phenomenon is well documented on the so-called small-world coefficient, correlation graphs of random system have small-world property. For this reason, alternative methods (including partial correlation) to graph construction were suggested. Using numerical simulations, we investigate other graph-theoretical properties, such as the rich-club coefficient and assortativity of correlation and partial correlation graphs. The results documents that both correlation and partial correlation graphs generated by a randomly coupled VAR(1) model show not only (spurious) small-world, but also other commonly discussed properties.

Keywords: assortativity, complex system, partial correlation, rich-club coefficient, VAR(1) process

Abstrakt. Pearsonova lineární korelace je nejčastějším přístupem ke konstrukci grafu konektivity daného komplexního systému. Bylo však ukázáno, že samotné použití korelace silně ovlivňuje odhadnutou strukturu původní sítě, a tedy i její grafově teoretické vlastnosti. Tento jev je podrobně zdokumentován především na takzvaném small-world koeficientu - korelační matice náhodného systému vykazuje vlastnost malého světa. Z tohoto důvodu byly navrženy alternativní metody konstrukce grafu konektivity jako například parciální korelace. Na základě numerických simulací určujeme další v praxi využívané grafově teoretické vlastnosti, jako jsou rich-club koeficient, a asortativita, korelačních a parciálních korelačních grafů. Ukazuje se, že korelační grafy, ale překvapivě i parciální korelační grafy generované vektorovým autoregresním procesem řádu 1 s náhodnou strukturní maticí vykazují nejen vlastnost malého světa, ale i výše uvedené vlastnosti.

Klíčová slova: asortativita, komplexní systém, parciální korelace, rich-club koeficient, VAR(1) proces

1 Introduction

The study of the structure of real dynamical (complex) systems is a subject of research in various scientific disciplines such as neuroscience, climatology, sociology, biology or

*This work has been supported by the grant No. SGS17/193/OHK4/3T/14 and by the Czech Science Foundation project No. GA19-16066S.

computer science. A key principle in complex network research is viewing the system at hand as a network of interacting subsystems (nodes), with one of the central questions being that of estimating the pattern of mutual or causal interactions of these. While in some cases (computer and social networks), the existence of links can be naturally defined, in other systems (neuroscience, climatology) it is often problematic to determine this structure by direct observation. In this context, the methodology of inference of functional (or effective) connectivity structure using only the knowledge of time series has recently been developed. Such networks are often described by global graph-theoretical characteristics, for example, clustering coefficient, characteristic path length, small-world property, efficiency, transitivity, rich-club coefficient, assortativity, and many others.

In practice, the Pearson linear correlation is the most commonly used approach to network construction. However, it has been shown that using the correlation matrix as the connectivity matrix of the system can affect resulting graph-theoretical properties. This phenomenon was described in detail especially for the small-world property (coefficient) [1]. Using correlation to graph construction leads, in particular with the knowledge of only a short sample of time series, to the false detection of the small-world property even at random graphs.

Alternative methods to the network construction which should mitigate this problem were suggested containing partial correlation, Granger causality or information-theoretical approach. In our work, we study the graph-theoretical properties, especially the rich-club coefficient and assortativity of the connectivity matrix in dependence on the form of graph construction (containing correlation and partial correlation).

2 Connectivity matrix of random process

A vector random process is a set of $\{\mathbf{X}_t |\ t \in M\}$, where $\mathbf{X}_t$ denotes n-dimensional vector random variable (n-dimensional vector whose components are random variables). The elements of vector random variables $\mathbf{X}_t$ will be denoted as X_t^i, where $i \in \hat{n}$, hence $\mathbf{X}_t = (X_t^1, \ldots, X_t^n)^\top$. Set M is usually represented as set of times, in cases where $M \subseteq \mathbb{N}$, or $M \subseteq \mathbb{Z}$ we talk about so-called time series or random process discrete in time. The motivation for representing a random process using a graph is to find out which elements of the system interact together. The set of these interactions between system elements is generally called connectivity. Thus, the first step in the graph representation is to assign the so-called connectivity matrix to the random process. The connectivity matrix is not a uniquely defined object, so there are different approaches to its definition. The most common approach is to define elements of this matrix using some measure of statistical dependence between elements of a given system, such as correlation or partial correlation. For simplicity, let us demonstrate the representation procedure with the choice of the correlation matrix as a matrix of connectivity. Alternative constructions of connectivity matrix, such as mentioned partial correlation, will be discussed later. Let us assume random variables X and Y, $E[X]$ denotes expected value of variable X and $\mathrm{var}(X)$ denotes its variance. Correlation coefficient of random variables X and Y is then defined by the expression

$$\rho(X,Y) = \frac{E\left[(X - E[X])(Y - E[Y])\right]}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}. \tag{1}$$

Correlation matrix of vector random variable $\mathbf{X}$, denoted as $\mathbf{C}(\mathbf{X})$, is matrix of correlation coefficients between all variables in the system, hence i,j-th element of correlation matrix is given by expression,

$$[\mathbf{C}(\mathbf{X}_t)]_{i,j} = \rho\left(X_t^i, X_t^j\right). \tag{2}$$

This correlation matrix $\mathbf{C}(\mathbf{X}_t)$ is then represented as the adjacency matrix of the graph.

Graph $G = (V, E)$ is ordered pair of sets where V is set of nodes and $E \subseteq \binom{V}{2}$ is set of edges. Elements of adjacency matrix of the graph G are defined as

$$a_{i,j} = \begin{cases} 1 & (i,j) \in E \\ 0 & (i,j) \notin E \end{cases}. \tag{3}$$

Since correlation matrix has real values, in order to be understood as an adjacency matrix, it is necessary to binarize it. In practice, threshold value K is chosen arbitrarily and adjacency matrix is constructed as follows,

$$[h_K(\mathbb{A})]_{i,j} = \begin{cases} 1 & a_{i,j} \geq K \\ 0 & a_{i,j} < K \end{cases}. \tag{4}$$

This thresholded correlation matrix (connectivity matrix of the original process) is then understood as adjacency matrix of the graph. On such constructed graph, graph-theoretical properties are then calculated.

Partial correlation

The partial correlation coefficient is designed to reflect the relationship between observed random variables, excluding the influence of other random variables in the system. Formally, partial correlation between random variables Y and Z with respect to variables $\mathbf{X} = (X^1, \ldots, X^n)^\top$ is defined as follows.

$\hat{Y} = \alpha_1 + \boldsymbol{\beta}_1^\top \mathbf{X}$ denotes best linear approximation of Y by variables $\mathbf{X}$ and $\hat{Z} = \alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X}$ denotes best linear approximation of Z by variables $\mathbf{X}$. Partial correlation between Y a Z with respect to $\mathbf{X}$ is defined as correlation of the residuals $Y - \hat{Y}$ and $Z - \hat{Z}$.

$$\rho(Y, Z|\mathbf{X}) = \rho\left(Y - \hat{Y}, Z - \hat{Z}\right). \tag{5}$$

Partial correlation matrix $\mathbf{P}$ of vector random variable $\mathbf{X} = (X^1, \ldots, X^n)^\top$ is defined as

$$[\mathbf{P}]_{i,j} = \rho\left(X^i, X^j | \mathbf{X} \setminus \{X^i, X^j\}\right). \tag{6}$$

In estimation we use the relation between partial correlation matrix and inverted correlation matrix. By $c_{i,j}$ is denoted the element of inverted correlation matrix, thus $[\mathbf{C}^{-1}(\mathbf{X})]_{i,j} = c_{i,j}$. Elements of partial correlation matrix $p_{i,j}$ are given by the expression

$$p_{i,j} = \begin{cases} -\frac{c_{i,j}}{\sqrt{c_{i,i} c_{j,j}}} & i \neq j \\ 1 & i = j \end{cases}. \tag{7}$$

In the process of characterization of random process as a graph, which was presented in this section, partial correlation coefficient (matrix) can be used instead of correlation coefficient (matrix).

3 Rich-club property

The first of the studied properties is rich-club property. This property (coefficient) was first introduced in paper studying Internet topology and it was designed to express whether high degree nodes are connected directly together [3]. In figure 1 you can see a typical example of rich-club structure. The rich-club coefficient $\phi(k)$ was defined as the ratio of the total actual number of links to the maximum possible number of links between elements with node degree at least k. Note that maximum possible number of these links is $N_{>k}(N_{>k}-1)/2$, where $N_{>k}$ is the number of nodes with degree greater than or equal to k. The rich-club coefficient of the graph is then defined by the expression

$$\phi(k) = \frac{2E_{>k}}{N_{>k}\left(N_{>k}-1\right)} \tag{8}$$

where $E_{>k}$ is the number of edges between the nodes of degree greater than or equal to k. If the value of $\phi(k)$ is close to 1 for values of k close to $k_{\max}$, the interpretation is that high degree nodes of the network are well connected and the graph has rich-club property. Value close to maximum node degree is not well defined, note that in our work we consider this number equal to $\lfloor 0.8k_{\max} \rfloor$. If we refer to the rich-club coefficient ϕ in the following text, we understand $\phi := \phi(\lfloor 0.8k_{\max} \rfloor)$.

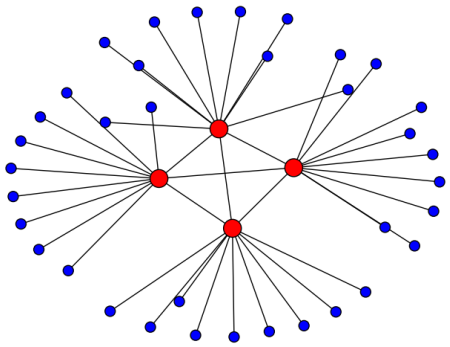


Figure 1: Example of "rich-club" network

Assortativity

A similar graph-theoretical property is assortativity. By assortativity, we understand a preference for graph nodes to attach to others that are similar in some way. The assortativity coefficient was defined as the Pearson correlation coefficient of degree between pairs of linked nodes [2]. Formally let us consider a network with n nodes and M edges with degree distribution p_k (probability that randomly chosen node has degree k). The original definition of assortativity deals with the so-called remaining degree (which is simply one less than node degree). Distribution of the remaining degree q_k is then given

by the expression

$$q_k = \frac{(k+1)\, p_{k+1}}{\sum_{j\geq 1} j p_j}, \tag{9}$$

where σ_q^2 is the variance of q_k is understood and e_{jk} denotes the joint probability distribution of the remaining degrees of the two nodes. Assortativity coefficient is finally defined as

$$r = \frac{\sum_{jk} jk\,(e_{jk} - q_j q_k)}{\sigma_q^2}, \tag{10}$$

which corresponds to correlation of degree between pairs of linked nodes, therefore its values are from $[-1, 1]$, for the random graphs $r = 0$ (for $n \to +\infty$), because edges are placed randomly without regard to node degree hence there is no correlation between node degrees.

4 Numerical simulation

In this section, we present the results of numerical simulations of the rich-club coefficient and assortativity of both correlation and partial correlation graphs. We consider systems with random interaction structure $\mathbb{A}$ which we model by Erdős-Rényi model of random graph (matrix). In this model the probability of presence of a direct link between each two elements is given by a predefined density value $D \in [0, 1]$. Practically we fix the required density of the matrix $\mathbb{A}$ (percentage of the direct links) and assign a value of 1 to the corresponding number of randomly selected elements. This binary matrix is further normalized to ascertain stationarity of the process by multiplying it with a constant $\frac{s}{\lambda_{\max}}$, where $\lambda_{\max}$ is the largest eigenvalue (in absolute value) of the matrix $\mathbb{A}$, and $s \in (0, 1)$ is an optional parameter. We set $s = 0.8$ throughout the paper. The testing dataset consists of vector autoregressive process of order 1 given by equation

$$\mathbf{X}_t = \mathbb{A}\mathbf{X}_{t-1} + \mathcal{E}_t. \tag{11}$$

as we said the above matrix $\mathbb{A}$ has a random structure. For this process, we can evaluate the covariance matrix (and hence also correlation matrix) analytically. For a symmetric matrix $\mathbb{A}$, the covariance matrix of VAR(1) process in the equation (11) is equal to equal to $(\mathbb{I} - \mathbb{A}^2)^{-1}$. Since matrix $\mathbb{A}$ is chosen randomly (process is representing random system) our expectation about values of assortativity coefficient of the (partial) correlation matrix is $r \approx 0$ - as we explained above, random system has assortativity coefficient equal to 0. In the case of rich-club coefficient ϕ we compute values of the rich-club coefficient of thresholded (partial) correlation matrix as well as rich-club coefficient of original structure matrix $\mathbb{A}$. This allows us to compare how much (partial) correlation affects the original structure in terms of the rich-club property rich-club property. Note that in this comparison of the rich-club coefficient, $k_{\max}$ is redefined as a minimum of maximal node degree of these two networks.

Numerical simulations were performed as follows. Random matrix $\mathbb{A}$ of dimension n was generated. Covariance matrix was calculated using expression $(\mathbb{I} - \mathbb{A}^2)^{-1}$ and by

normalization we obtain correlation matrix . From this correlation matrix, partial correlation matrix was evaluated using (7). These matrices were then thresholded to the density D. These thresholded matrices served us as adjacency matrices of the original system (process). From these matrices, rich-club property and assortativity were evaluated. Below we present results of rich-club coefficient and assortativity of correlation and partial correlation graphs generated by VAR(1) process with random structural matrix with various values of network size n and density D.

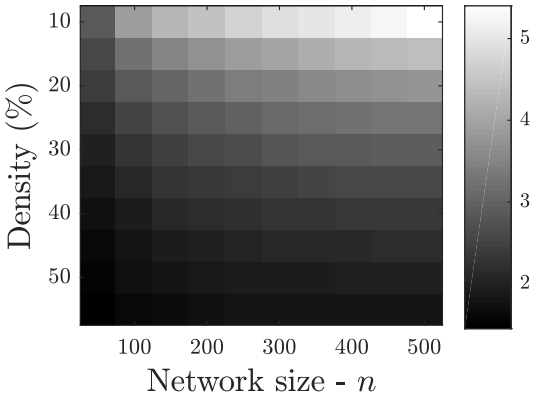


Figure 2: Ratio of the rich-club coefficient of thresholded correlation matrix and rich-club coefficient of the original structural matrix $\mathbb{A}$. Results are shown for a range of matrix sizes: $n \in \{50, 100, \ldots, 500\}$ and threshold densities $D \in \{10, 15, \ldots, 55\%\}$.

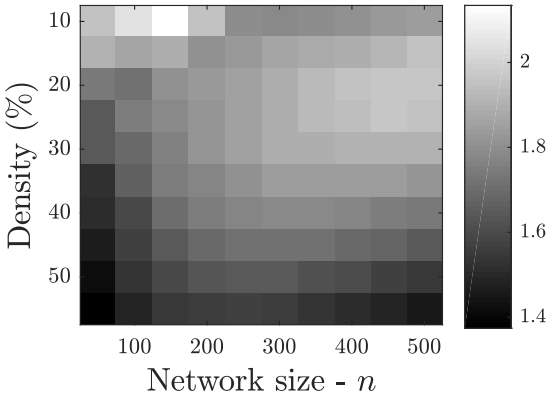


Figure 3: Ratio of the rich-club coefficient of thresholded partial correlation matrix and rich-club coefficient of the original structural matrix $\mathbb{A}$. Results are shown for a range of matrix sizes: $n \in \{50, 100, \ldots, 500\}$ and threshold densities $D \in \{10, 15, \ldots, 55\%\}$.

As can be seen from figures 2 and 3 values of ratio of the rich-club coefficient of (partial) correlation graphs and rich-club coefficient of random structural matrix $\mathbb{A}$ are greater than 1. This observation can be interpreted as the using of correlation or partial correlation itself increases the observed rich-club coefficient of the investigating system. In both cases, the biggest value of this ratio is acquired for large and relatively sparse

networks. In the case of correlation graphs, ratio reaches the maximum value of 5, in partial correlation graphs approximately 2. Thus partial correlation is biasing the resulting network less than correlation but still causes a relatively strong rich-club structure. Note that even for relatively small and dense networks is this ratio greater than 1 and so bias is seen in both cases.

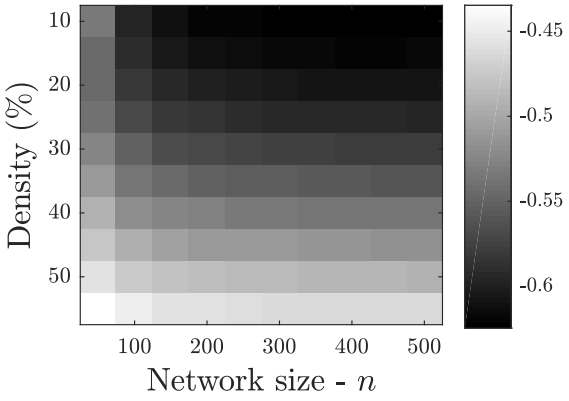


Figure 4: Assortativity coefficient r of thresholded correlation matrix. Results are shown for a range of matrix sizes: $n \in \{50, 100, \ldots, 500\}$ and threshold densities $D \in \{10, 15, \ldots, 55\%\}$.

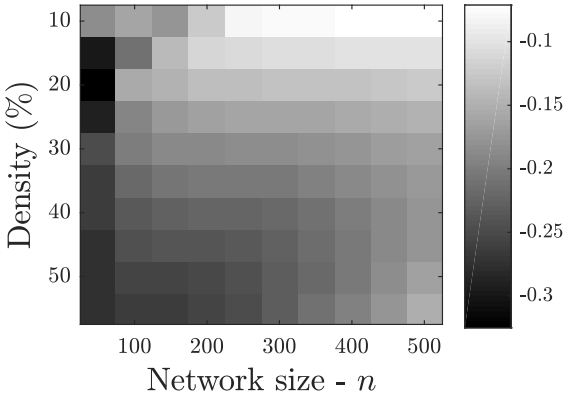


Figure 5: Assortativity coefficient r of thresholded partial correlation matrix. Results are shown for a range of matrix sizes: $n \in \{50, 100, \ldots, 500\}$ and threshold densities $D \in \{10, 15, \ldots, 55\%\}$.

Figure 4 suggests that correlation graphs show low values of the assortativity coefficient thus correlation networks are relatively disassortative. In the case of partial correlation graphs, see figure 5, this phenomenon is not as strong but also there the original random structure, for which is characteristic $r \approx 0$, is wiped out.

5 Conclusion

The main aim of this paper was to study the influence of using correlation and partial correlation in the process of representing the system by graph to the selected graph-theoretical properties: rich-club property and assortativity. Our simulations demonstrated that even in completely random systems these structures or properties can be found just because of the using of these methods. In practice (neuroscience, sociology, climatology, economics,...) the above procedure of representing a real dynamical system by a graph is widely applied and based on measured graph-theoretical properties conclusions of the original system are drawn. As this paper or [1] illustrates, these conclusions can be misleading because even in random systems internal structures like small-world property, rich-club property or assortativity can be found just because of the methodology of representing. For this reason, we recommend using some kind of effective connectivity method, for example, Granger causality or methods based on transfer entropy. In these methods, the resulting matrix at least converges to the original structural matrix of the system and graph-theoretical properties should not be biased.

References

[1] J. Hlinka, D. Hartman, N. Jajcay, D. Tomecek, J. Tintera, M. Palus. *Small-world bias of correlation networks: From brain to climate*. Chaos **27** (2017)

[2] M. Newman. *Assortative Mixing in Networks* Physical Review Letters **89**(20) (2012)

[3] S. Zhou, J. Mondrago. *The Rich-Club Phenomenon In The Internet Topology* . IEEE Communications Letters **8**(3) (2004)

Image Invariants to Anisotropic Gaussian Blur*

Jitka Kostková

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

školitel: Jan Flusser, Ústav teorie informace a automatizace AV ČR, v.v.i.

Abstract. The paper presents a new theory of invariants with respect to Gaussian blur. Unlike earlier methods, the blur kernel may be arbitrary oriented, scaled and elongated. Such blurring is a semi-group action in the image space, where the orbits are classes of blur-equivalent images. Every orbit is represented by a *primordial image* ("maximally deconvolved image"). We propose a non-linear projection operator P which extracts blur-insensitive component of the image. Projection operator P divides each image f into its Gaussian component Pf (projection onto the set of all Gaussian functions) and a non-Gaussian component. It can be proven that

$$I(f) = \frac{\mathcal{F}(f)}{\mathcal{F}(Pf)}$$

is an invariant w.r.t. Gaussian blur in the frequency domain. The invariants in the image domain are then formally defined as moments of the blur-insensitive component but can be computed directly from the blurred image without an explicit construction of the projections

$$M_{pq} = \frac{m_{pq}^{(f)}}{m_{00}} - \sum_{\substack{l=0 \\ l+k\neq 0, \\ l+k \text{ even}}}^{p} \sum_{k=0}^{q} \binom{p}{l}\binom{q}{k} \sum_{\substack{i=0 \\ j\geq \frac{k-l}{2}}}^{\lfloor \frac{k}{2} \rfloor} \sum_{j=0}^{i} (-1)^{i-j} \binom{k}{2i}\binom{i}{j} (l+k-2i-1)!! \cdot$$
$$\cdot (2i-1)!! \left(\frac{m_{11}}{m_{00}}\right)^{k-2j} \left(\frac{m_{20}}{m_{00}}\right)^{\frac{l-k}{2}+j} \left(\frac{m_{02}}{m_{00}}\right)^{j} M_{p-l,q-k}\,.$$

Image description by the new invariants does not require any prior knowledge of the particular blur kernel shape and does not include any deconvolution. Potential applications are in blur-invariant image recognition and in robust template matching.

Keywords: Gaussian blur, Semi-group, Projection operator, Image moments, Moment invariants

Abstrakt. V tomto článku prezentujeme novou teorii invariantů vůči gaussovskému rozmazání. Narozdíl od existujících metod uvažujeme konvoluční jádra, která jsou libovolně natočená, protažená a škálovaná. Tato rozmazání jsou akce semigrupy na prostoru obrazových funkcí, kde orbity představují třídy ekvivalentních obrázků. Každá orbita je reprezentována tzv. *praobrázkem*, který si lze představit jako „maximálně dekonvolvovaný" obrázek. Navrhli jsme projekční operátor P, který extrahuje komponentu obrázku necitlivou na rozmazání. Projekční operátor

*This work was supported by the Czech Science Foundation (Grant No. GA18-07247S), by the *Praemium Academiae*, and by the Grant Agency of the Czech Technical University (Grant No. SGS18/188/OHK4/3T/14).

P rozdělí obrázek f na gaussovskou komponentu Pf (projekce na množinu všech gaussovských funkcí) a negaussovskou komponentu. Lze ukázat, že

$$I(f) = \frac{\mathcal{F}(f)}{\mathcal{F}(Pf)}$$

je invariant vůči gaussovskému rozmazání ve frekvenční oblasti. Invarianty v obrazové oblasti lze formálně definovat jako momenty praobrázku, ale lze je spočítat přímo z rozmazaného obrázku, aniž bychom explicitně konstruovali projekce

$$M_{pq} = \frac{m_{pq}^{(f)}}{m_{00}} - \sum_{\substack{l=0 \\ l+k\neq 0, \\ l+k \text{ even}}}^{p} \sum_{k=0}^{q} \binom{p}{l} \binom{q}{k} \sum_{\substack{i=0 \\ j\geq \frac{k-l}{2}}}^{\lfloor \frac{k}{2} \rfloor} \sum_{j=0}^{i} (-1)^{i-j} \binom{k}{2i} \binom{i}{j} (l+k-2i-1)!! \cdot$$
$$\cdot (2i-1)!! \left(\frac{m_{11}}{m_{00}}\right)^{k-2j} \left(\frac{m_{20}}{m_{00}}\right)^{\frac{l-k}{2}+j} \left(\frac{m_{02}}{m_{00}}\right)^{j} M_{p-l,q-k} \,.$$

Metoda nevyžaduje apriorní znalost konkrétního konvolučního jádra a nezahrnuje dekonvoluci. Tyto deskriptory lze využít pro rozpoznávání obrazu necitlivé na rozmazání a robustní databázové vyhledávání.

Klíčová slova: gaussovské rozmazání, semigrupa, projekční operátor, obrazové momenty, momentové invarianty

Full paper: J. Kostková, J. Flusser, M. Lébl and M. Pedone. *Image Invariants to Anisotropic Gaussian Blur*. Proceedings of the Scandinavian Conference on Image Analysis – SCIA'19, Norrköping, 11–13 June 2019, 140–151.

Linear Classification on Top Samples

Václav Mácha

3rd year of PGS, email: machava2@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Václav Šmídl, Department of Adaptive Systems
Institute of Information Theory and Automation, CAS
Lukáš Adam, Department of Computer Science and Engineering
Southern University of Science and Technology, Shenzhen, China

Abstract. Many interesting binary classification problems minimizes misclassification below (or above) a decision threshold. Such problems include, for example, Top Push [3], Accuracy at the Top [1, 2] or the problem of hypothesis testing [4]. Regardless of the similarity of these tasks, they are all considered separate problems. In this work, In this work, we show that all of them can be written in the same form as well as we propose a general framework to handle these classes of problems. Moreover, we show which known methods fall into the framework and derive new methods or improvements of known methods, which also fall into this framework too. Besides that, we provide a theoretical analysis, where we show that all presented methods can be transformed or at least approximated by a convex optimization problem. Furthermore, we mention selected possible pitfalls the methods may encounter. We also suggest several numerical improvements including the implicit derivative and stochastic gradient descent. The last part of the work contains an extensive numerical study of all mentioned methods. Based both on the theoretical properties and numerical experiments, we conclude the paper by suggesting which method should be used in which situation.

Keywords: Binary Classification, Accuracy at the Top, Neyman-Pearson, Top-Push, Convexity.

Abstrakt. Mnoho zajímavých binárních klasifikačních problémů minimalizuje počet špatně klasifikovaných pozorování pod (nebo nad) rozhodovacím prahem. Mezi takové problémy patří například úloha Top Push [3], Accuracy at the Top [1, 2] nebo úloha testování hypotéz [4]. Všechny tyto problémy jsou uvažovány odděleně bez ohledu na jejich podobnost. V této práci ukazujeme, že všechny lze zapsat stejným způsobem a také navrhujeme obecný rámec pro řešení úloh spadajících to těchto tříd problémů. Dále ukazujeme, které známe metody spadají do obecného rámce a odvozujeme nové metody nebo vylepšení metod známých, která také spadají do uvedeného obecného rámce. Kromě toho poskytujeme teoretickou analýzu tohoto rámce, kde ukazujeme, že všechny uvedené metody lze zapsat nebo alespoň aproximovat pomocí konvexní optimalizační úlohy. Dále zmiňujeme možná úskalí, která mohou pro některé metody nastat. Mimoto navrhujeme několik numerických vylepšení včetně užití implicitních derivací a metody stochastického gradientního sestupu. Poslední část práce obsahuje rozsáhlou numerickou studii uvedených metod. Na základě teoretických vlastností i numerických experimentů zakončujeme práci doporučením, pro jaké úlohy použít jakou metodu.

Klíčová slova: Binární klasifikace, Accuracy at the Top, Neyman-Pearson, Top-Push, Konvexita.

Full paper: L. Adam, V. Mácha, V. Šmídl, T. Pevný. *General Framework for Bi-*

nary Classification on Top Samples. Submitted to Journal of Machine Learning Research (2019).

References

[1] S. Boyd, C. Cortes, M. Mohri, and A. Radovanovic. *Accuracy at the top.* In Advances in neural information processing systems (2012), 953–961.

[2] M. Grill and T. Pevný. *Learning combination of anomaly detectors for security domain.* Computer Networks **107** (2016), 55–63.

[3] N. Li, R. Jin, and Z.-H. Zhou. *Top rank optimization in linear time.* In Advances in neural information processing systems, NIPS'14, MIT Press (2014), 1502–1510.

[4] J. Neyman and E. S. Pearson. *IX. On the problem of the most efficient tests of statistical hypotheses.* Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **231** (1933), 289–337.

Maniputlation of Time Evolution in Quantum Purification Protocol*

Martin Malachov

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

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

Abstract. One of proposed quantum protocol allowing entanglement purification has been shown to exhibit chaotic properties. We numerically analyse much wider class of generalised protocols showing how the new parameters manifest in the chaotic features. One of the parameters can be understood as free time evolution added to protocol iterations and we show that this time relaxation can be manipulated by other protocol parameters. This result allows e.g. to virtually introduce or eliminate the free evolution leading to modify chaotic evolution represented by fractal structures. Furthermore in this sense of time manipulation we can execute a whole new family of protocols by a single protocol setting. In consequence a whole family of protocols is redundant to further deeper study.

Keywords: qubit, chaos, time evolution

Abstrakt. Jeden z navržených purifikačních protokolů k purifikaci kvantového provázání vykazuje chaotické vlastnosti. Nyní studujeme celou novou třídu obecnějších protokolů a ukazujeme, jak nově zavedené parametry ovlivňují chaotické efekty. Jeden z parametrů může být chápán jako přidání volného časového vývoje mezi iterace protokolu; ukážeme, že tato časová relaxace může být ovládána otatními parametry protokolu. To vede na příklad k tomu, že lze virtuálně zavést či naopak vyrušit časový vývoj v protokolu, což modifikuje chaotické chování ve formě fraktálních struktur. Navíc v tomto smyslu můžeme celou skupinu protokolů realizovat jediným protokolem. V důsledku tak je možné z dalších studií protokolů jeden z parametrů vynechat.

Klíčová slova: qubit, chaos, časový vývoj

1 Introduction

With quantum technologies getting more and more powerful, eyes turn to a fundamental question. Realistic quantum state suffers from decoherence because it interacts with its environment and this leads e.g. to dissipation of entanglement which is an important resource for quantum computation. Is there a way to effectively protect the information or repair it? One of quantum protocols [1] proposed twenty years ago suggests to use measurement based selection and modification to protect the maximally entangled Bell state. This nonlinear protocol was later shown to exhibit chaotic behaviour [4]. This chaos is bound to the iterative evolution of quantum states itself and so has no analogy

*This work has been supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS19/186//OHK4/3T/14.

in classical physics. The chaos manifests in forming fractal structure [2] of chaotic states. In addition, when evolution of mixed states is studied we come to a particular behaviour understood as a phase transition where the physical phase is formed by the induced fractal structure and the purity plays the role of temperature, [5]. There are many interesting questions to the open quantum systems evolution and we hope to uncover some of the mysteries of general quantum physics.

In this work we present a generalisation of the original protocol. The motivation to this can be found in desire for general understanding of nonlinear evolution of quantum systems. New protocols could possess interesting properties, allow for more effective entanglement purification etc. Our finding brings important results thanks to particular parameterisation of the protocol action and numerical tools. One of the important thought is to understand general protocol as another protocol equipped with additional time relaxation operator. This operator we interpret as a free time evolution inserted between subsequent protocol iterations. The question of delay between protocol iterations is fundamental for practical applicability of the protocol. Besides the time phase determined by this time relaxation, the protocol is determined by two real parameters and such general protocol acts on a mixed qubit state which can be described using three real parameters. This number of parameters together with nonlinear evolution equations turns the simple idea of the protocol into formidable task to study despite the model is one of easiest "toy models" one can consider.

2 Protocol settings

Original motivation for the nonlinear protocol lies in the fact that closed systems evolve under unitary transformation which is isometry. In such case a general quantum state cannot be repaired if damaged by environment. A simple idea implemented by Bechmann-Pasquinucci [1] suggests to use another copy of the system as a particular environment and based on the measurement of this environment to decide to repair the desired information with suitable unitary gate. [1] gives simple scheme which preserves Bell state $|\Phi^+\rangle$ and later papers show [4, 5] that the protocol is truly capable of repairing the damaged $|\Phi^+\rangle$, i.e. purifying the entanglement. The protocol uses two key elements: CNOT gate to implement measurement-based selection and Hadamard gate. While the CNOT gate is responsible for nonlinear evolution, the Hadamard gate can be replaced by general unitary gate U as so called twirling operator. Therefore we now introduce modified protocol with scheme showed at 1.

3 Protocol modification

For the sake of simplicity we consider only action of the protocol on a single qubit, not qubit pairs as the original protocol [1, 4]. This step will dismiss any entanglement out of discussions but it allows us to easily perform numerical analysis which is the only tool available for this mathematical problem. The most general form of qubit is mixed state ρ characterised by three real parameters in following way, often called (Bloch-)Fano

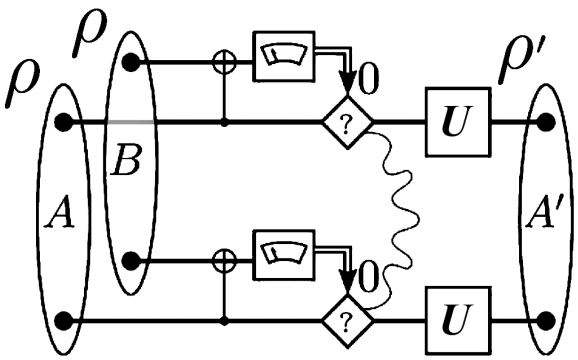


Figure 1: Scheme of a general protocol. Two copies of qubit pairs undergo measurement based modification which depends on twirling gate U.

representation, [3]:

$$\rho = \frac{1}{2}\begin{pmatrix} 1+w & u-iv \\ u+iv & 1-w \end{pmatrix};\ u,v,w \in \mathbb{R} \wedge u^2+v^2+w^2 \leq 1 \tag{1}$$

The Fano space is a ball in a threedimensional Cartesian space and the north/south poles $u = v = 0, w = \pm 1$ are the basis states of given computational basis.

The so called twirling gate U is now arbitrary unitary gate and it is expected to modify chaotic features of the protocol. The gate U can be expressed as a matrix (in the computational basis) depending only on three real parameters, angles. Fourth parameter needed to uniquely describe a unitary matrix only influences global phase of the physical state and as such is redundant. Additionally, the matrix can be decomposed in following way:

$$U = T_\tau R_{x,\psi};\quad T_\tau = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\tau} \end{pmatrix},\ R_{x,\psi} = \begin{pmatrix} \cos x & \sin x e^{i\psi} \\ -\sin x e^{-i\psi} & \cos x \end{pmatrix}. \tag{2}$$

The reason for such decomposition is that matrix T_τ can be given physical interpretation of free time evolution of the system. Such free time evolution manifests on basis states as a gain of a relative "time" phase τ. The protocol equipped with U is physically understood as a protocol equipped with rotation R followed by free time evolution, i.e. iterations of the protocol are separated by certain time interval. When one iteration is executed on state ρ, it is changed as follows:

$$\rho \to \rho' = \frac{U \cdot (\rho \odot \rho) \cdot U^\dagger}{\mathrm{Tr}(U \cdot (\rho \odot \rho) \cdot U^\dagger)} \tag{3}$$

where symbol $\odot$ stands for elementwise multiplication of matrices. Considering Fano representation of the state ρ we can write evolution equations in u, v, w:

$$u' = \frac{-2w\sin 2x\cos(\psi-\tau) + (u^2-v^2)\left[\sin\psi\sin(\psi-\tau) + \cos 2x\cos\psi\cos(\psi-\tau)\right] + 2uv\left[\cos\psi\sin(\psi-\tau) - \cos 2x\sin\psi\cos(\psi-\tau)\right]}{1+w^2},$$
$$v' = \frac{2w\sin 2x\sin(\psi-\tau) + (u^2-v^2)\left[\sin\psi\cos(\psi-\tau) - \cos 2x\cos\psi\sin(\psi-\tau)\right] + 2uv\left[\cos\psi\cos(\psi-\tau) + \cos 2x\sin\psi\sin(\psi-\tau)\right]}{1+w^2},$$
$$w' = \frac{2w\cos 2x + (u^2-v^2)\cos\psi\sin 2x - 2uv\sin\psi\sin 2x}{1+w^2}. \tag{4}$$

These nonlinear equations are quadratic rational in state coordinates u, v, w; terms $2w, u^2-v^2, 2uv$ are consequence of the CNOT gate application. These terms are geometrically combined by functions of protocol angles x, ψ, τ. Note that the combinations have property $\alpha \cdot 2w + \beta \cdot (u^2-v^2) + \gamma \cdot 2uv \Rightarrow \alpha^2+\beta^2+\gamma^2 = 1$ for all new coordinates u', v', w'.

4 Asymptotic dynamics

We are interested in asymptotic dynamics of the protocol because of the chaotic features. Fractal structures emerge with repetitive iterations of the protocol and as such are demanding on computational power, they are simulated by sufficient amount of iterations. There is no other approach besides numerical because there is no mathematical background for equations 4. Problem of chaos with three real variables goes beyond scope of mathematical books and there are no usable analytical statements.

The two connected questions we try to answer at the moment are: What fractal structures are obtainable for general protocol? What attractors are determined by the protocol? Naturally, the answer is hidden in values of parameters x, ψ, τ. So we must ask how do the fractal structures and attractors depend on the protocol parameters.

Application of one iteration of protocol equipped with $U = TR$ twirling gate will be schematically written in following notation:

$$\rho \to \rho' = TR[\rho] \tag{5}$$

To give visual representation to calculations and conclusions of $TR[TR[\dots TR[\rho]\dots]]$ we now discuss creation of images we call *attractor maps*. Such maps serve to identify attractor of given initial state. The map can be computationally a matrix where position (matrix element position=pixel) represents an initial state and matrix element value would code the attractor. Such matrix can be visualised as a coloured rectangular image where the matrix entries are converted to colours with some colour scheme. The two-dimensional rectangle must be now confronted with the ball of all possible initial states, 1. We use following physically motivated approach: choosing value of so called purity $\mathcal{P} = u^2 + v^2 + w^2$ we choose a sphere of certain radius. In this two-dimensional manifold the initial state is than uniquely determined by two spherical angles φ, ϑ. The attractor maps will code value φ on x-axis and ϑ on y-axis. Attractors will be coded in grayscale. Such coding would not be generally unique but for our purposes it recognises individual attractors which is the aim of attractor maps. Connected regions of the same colour can be understood in analogy to basins of attraction introduced in theory of complex functions [6]. Borders separating regions of different colours form the structure which is of our interest as it contains states sensitive to initial conditions. That is the main point crucially connected to chaos. The aim of this paper is not to analyse the fractals but to give a brief overview of relation between the fractal and protocol parameters.

Consider two protocol characterised with two sets $x_1, \psi_1, \tau_1,\ x_2, \psi_2, \tau_2$. We will mark corresponding iterations with $T_1R_1[.], T_2R_2[.]$. We can say two protocols are *asymptotically equivalent* in case they induce the same fractal structure in the asymptotic regime. The fractal structure may be rotated (as the ball in the Fano representation, 1) and the attractors may be different but the crucial properties of the fractal structure like its dimension remain kept. This concept of asymptotic equivalence define a class of protocols which in attractor maps differ only in shift of position and change of colours, but "the oceans and continents" keep their shape.

Numerical analysis uncovers peculiar result: Two protocols $T_1R_1[.], T_2R_2[.]$ are asymptotically equivalent if $x_1 = x_2 \wedge \psi_1 - 2\tau_1 = \psi_2 - 2\tau_2$. We now define angle $\Delta := \psi - 2\tau$ for given protocol and we conclude that the induced fractal structure of the protocol depends

only on two real numbers: x, Δ. Before we proceed to practical applications we give also a stronger result. We write it symbolically:

$$T_1R_1[T_1R_1[T_1R_1[\dots T_1R_1[\rho]\dots]]] = T_0^\dagger\cdot(T_2R_2[T_2R_2[T_2R_2[\dots T_2R_2[T_0\cdot\rho\cdot T_0^\dagger]\dots]]])\cdot T_0. \quad (6)$$

This formula is independent of number of iterations. That means not only two protocol are equivalent in asymptotic regime but they can be put equal at a price of additional Fano space transformation expressed by operators T_0. This operator is used before the protocol application and its inverse is applied after the protocol is executed. Symbol T_0 is used because the operator has the form of free time evolution 2 and corresponding phase is found: $\tau_0 = \tau_1 - \tau_2$. The action of operator T_0 on states in Fano space manifests as a rotation around w-axis for angle τ_0.

When given two asymptotically equivalent protocols x, Δ, the value τ_0 explains: Perform iterations of the first protocol. You will obtain the same result as if you would rotate the Fano space for angle τ_0, perform the iterations of the other protocol and rotated the Fano space back by $-\tau_0$. In the images of attractor maps the first rotation shifts the fractal structure by certain angle to the left (rotation of the Fano space around w-axis is the shift in the spherical angle φ which is on the horizontal axis of the map). The later rotation changes the attractors which results into change of the colours of the map.

Two equivalent protocols do not have to be studied individually. The fractal structure keeps the same properties and the fact we know the additional transformation T_0 allows us to easily find attractors of one of equivalent protocols by simply rotating attractors of another one for corresponding angle τ_0. We conclude that regarding asymptotic analysis of general protocol, only two parameters x, Δ are relevant while τ_0 allows to specify exact attractor of given state ρ. The fractal structure efficiently depends on only two parameters, see example of attractor maps in 2.

Now, we proceed to applications of previous result which we put in physical context. The time evolution expressed with phase τ can be handled using equation 6 at the cost of setting appropriate ψ. We present three basic concepts of this time manipulation.

- *Adding the time relaxation to the protocol.*

  $$R_1[R_1[R_1[\dots R_1[\rho]\dots]]] \to T\cdot(TR_2[TR_2[TR_2[\dots TR_2[T^\dagger\cdot\rho\cdot T]\dots]]])\cdot T^\dagger \quad (7)$$

  If the protocol iterations are executed immediately but we need certain time evolution, resp. the phase τ to take place, we can implement it without actually leaving the system evolve freely. The required time for protocol implementation can dramatically decrease. Gate R_2 must be chosen with $\psi_2 = \psi_1 + 2\tau$.

- *Eliminating the time relaxation from the protocol.*

  $$TR_1[TR_1[TR_1[\dots TR_1[\rho]\dots]]] \to T^\dagger\cdot(R_2[R_2[R_2[\dots R_2[T\cdot\rho\cdot T^\dagger]\dots]]])\cdot T \quad (8)$$

  On contrary, if the protocol iterations are separated by time evolution (for example due to transmission of state or due to technological realisation) we can modify the protocol to undo this unwanted phase τ with setting $\psi_2 = \psi_1 - 2\tau$.

- *Changing the time relaxation of the protocol.*

  The most general case is exactly the equation 6 with meaning that the time delay between protocol iterations can be lengthened or shortened at will.

All cases require pre- and postmodification of the state to evolve. The modification is free evolution for a specific time delay which is determined by the default and the desired protocol.

5 Conclusion

We have presented results based on numerical evidence but with high relevance and reliability. These very interesting statements can possibly be put on more solid ground by certain analytical calculations. However, we deal with infinite iterations of multidimensional nonlinear maps which go beyond the scope of available mathematical books. Therefore, the numerical approach is very welcome to at least give intuition what features and consequences might the general protocol possess.

The main result 6 is valid for any number of iterations and allows us to identify one-parametrical classes of protocols with the same x, Δ. The protocols belonging to the same class induce the same fractal structure in the Fano space. The difference between the protocols is given by actual value of τ (resp. $\psi = \Delta + 2\tau$). This parameter then rotates the fractal structure and attractors as illustrated in 2.

Our findings can be used in several ways, most of all to modify free time evolution between protocol iterations. The price for this is in using additional unitary transformation before and after the protocol iterations. We have presented exact forms of these operators.

The importance of our results also lies in deeper understanding of nonlinear protocols and further simplification of their study.

References

[1] H. Bechmann-Pasquinucci et al. Phys. Lett. A **242** (1998), 198–204

[2] K. Falconer. *Fractal Geometry, Mathematical Foundations and Applications*. Wiley, New York (1990)

[3] U. Fano. Rev. Mod. Phys. **29** (1957), 74–93.

[4] T. Kiss, S. Vymětal, L.D. Tóth, A. Gábris, I. Jex, G. Alber. Phys. Rev. Lett. **107** (2011), 100501

[5] M. Malachov, I. Jex, O. Kálmán, T. Kiss. Chaos **29** (2019), 033107

[6] J.W. Milnor. *Dynamics in One Complex Variable*, 3rd edition. Princeton University Press (2000)

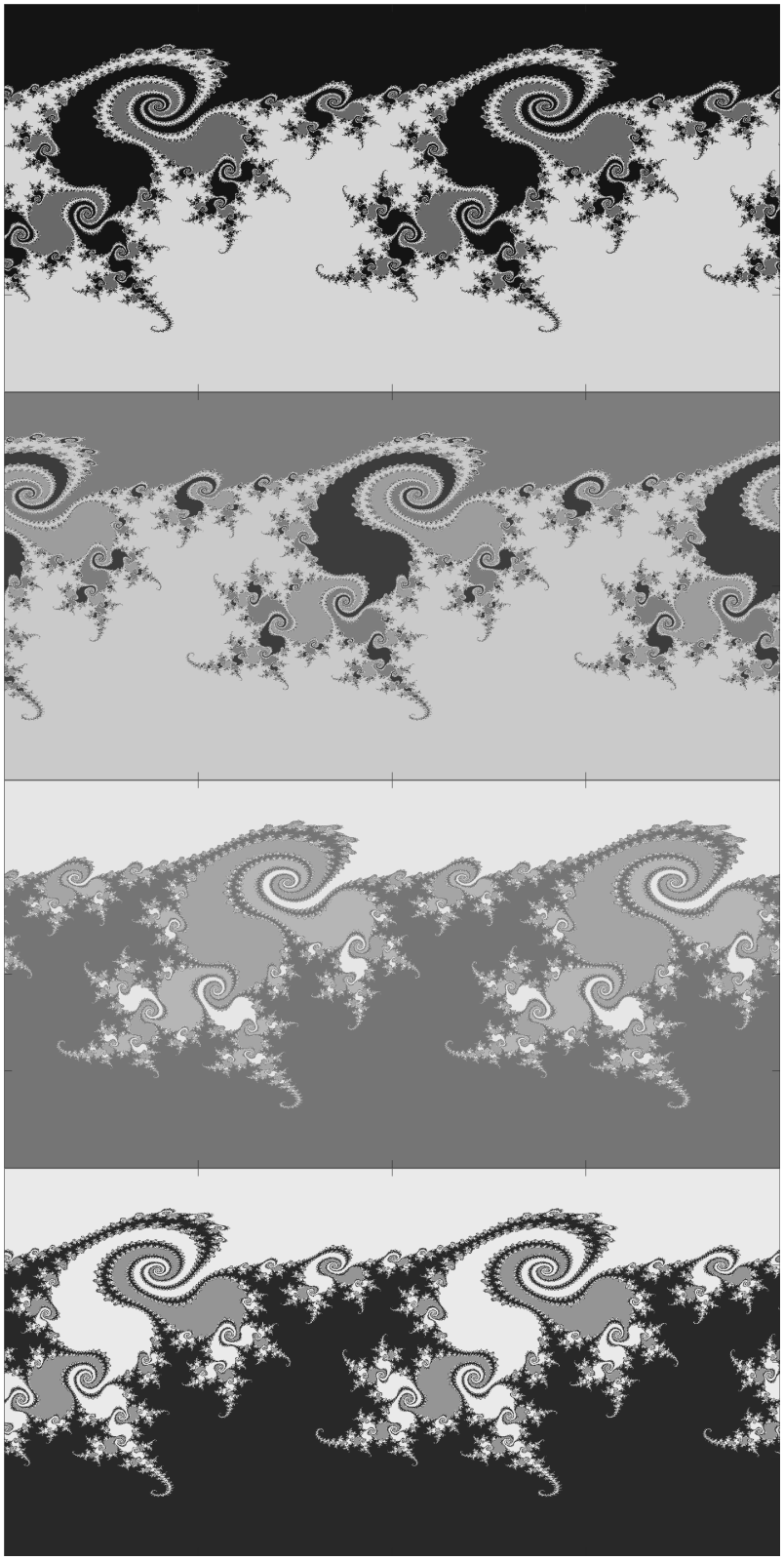


Figure 2: The attractor maps of protocols with $x = 30°, \Delta = 89°$, from top to bottom τ is chosen $0°, 60°, 120°, 180°$. These maps depict asymptotic evolution of pure states, the Bloch sphere is translated to rectangular maps via spherical angles $\varphi \in [0, 2\pi), \vartheta \in [0, \pi]$ which are assigned to horizontal and vertical axes. Each colour codes a different attractor, i.e. initial states (pixels) with the same colour converge to the same attractor state. Choice of higher τ changes colours and shifts the fractal structure to the left.

Derived Sequences of Arnoux–Rauzy Sequences*

Kateřina Medková

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

advisors:

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

Karel Klouda, Department of Applied Mathematics
Faculty of Information Technology, CTU in Prague

Abstract. One of the key notions in combinatorics on words are derived sequences which were introduced by Durand [1]. Let $\mathbf{u} = u_0u_1u_2\cdots$ be an infinite sequence with all letters u_i from a finite alphabet $\mathcal{A}$ and let $w = u_iu_{i+1}\cdots u_{i+n-1}$ be a non-empty factor of $\mathbf{u}$. The integer i is called an *occurrence* of the factor w in $\mathbf{u}$. Let $i < j$ be two consecutive occurrences of w in $\mathbf{u}$. Then the word $u_iu_{i+1}\cdots u_{j-1}$ is a *return word* to a factor w in $\mathbf{u}$. Here we take into consideration only the sequence $\mathbf{u}$ whose each factor w has finitely many return words. Such a sequence is called *uniformly recurrent*. In addition, if w is a prefix of $\mathbf{u}$, then the sequence $\mathbf{u}$ can be written as the unique concatenation of the return words to w. The ordering of the return words in this concatenation is coded by the *derived sequence* of $\mathbf{u}$ to its prefix w.

Return words and derived sequences were especially studied in the case of Sturmian sequences, which are the aperiodic binary sequences having the least factor complexity possible. In particular, Vuillon [4] showed that a binary sequence is Sturmian if and only if its each factor has exactly two return words. This property implies that the derived sequence to each prefix of a Sturmian sequence is Sturmian as well.

Sturmian sequences have various generalizations for multi-literal alphabets, one of them are Arnoux–Rauzy sequences. A uniformly recurrent sequence $\mathbf{u}$ over $\mathcal{A}$ is called *Arnoux–Rauzy* if for all n it has $(\#\mathcal{A}-1)n+1$ factors of length n with exactly one left and one right special factor of length n. In [2] the authors show that each factor of an Arnoux–Rauzy sequence over $\mathcal{A}$ has exactly $\#\mathcal{A}$ return words. It means that derived sequences of Arnoux–Rauzy sequences over $\mathcal{A}$ can be considered over the same alphabet $\mathcal{A}$. Nevertheless, such a property does not characterize Arnoux–Rauzy sequences if $\#\mathcal{A} > 2$.

The aim of this paper is to study derived sequences of Arnoux–Rauzy sequences. For every Arnoux–Rauzy sequence $\mathbf{u}$ we describe the set $\mathrm{Der}(\mathbf{u})$ of derived sequences to all non-empty prefixes of $\mathbf{u}$. As a corollary, we show that every derived sequence of an Arnoux–Rauzy sequence is an Arnoux–Rauzy sequence as well. Durand [1] proved that the sequence $\mathbf{u}$ is primitive substitutive, i.e. it is a morphic image of a fixed point of a primitive morphism, if and only if the set $\mathrm{Der}(\mathbf{u})$ is finite. Here we precisely determine the cardinality of $\mathrm{Der}(\mathbf{u})$ for all primitive substitutive Arnoux–Rauzy sequences. It generalizes the results from [3], where the cardinality of $\mathrm{Der}(\mathbf{u})$ is bounded for the fixed points of primitive Sturmian morphisms.

*This work has been supported by the project no. CZ.02.1.01/0.0/0.0/16_019/0000778 and by the grant SGS17/193/OHK4/3T/14.

Keywords: Arnoux–Rauzy sequence, derived sequence, return word

Abstrakt. Derivovaná slova definoval Durand [1] v roce 1998 a od té doby se stala jedním z klíčových pojmů kombinatoriky na slovech. Nechť $\mathbf{u} = u_0u_1u_2\cdots$ je nekonečné slovo nad konečnou abecedou $\mathcal{A}$ a nechť $w = u_iu_{i+1}\cdots u_{i+n-1}$ je jeho neprázdný faktor. Přirozené číslo i nazýváme *výskytem* faktoru w ve slově $\mathbf{u}$. *Návratové slovo* k faktoru w je slovo $u_iu_{i+1}\cdots u_{j-1}$, kde $i < j$ jsou dva po sobě jsoucí výskyty w ve slově $\mathbf{u}$. V tomto článku uvažujeme pouze nekonečná slova, jejichž každý faktor má konečně mnoho návratových slov. Taková slova nazýváme *uniformně rekurentní* a platí pro ně, že pro libovolný prefix w slova $\mathbf{u}$ můžeme slovo $\mathbf{u}$ jednoznačně zapsat jako zřetězení návratových slov k prefixu w. Pořadí jednotlivých návratových slov v tomto zřetězení je kódováno *derivovaným slovem* slova $\mathbf{u}$ k prefixu w.

Návratová a derivovaná slova jsou detailně prostudována zejména v případě sturmovských slov, což jsou binární aperiodická slova s minimální faktorovou komplexitou. Vuillon [4] ukázal, že binární nekonečné slovo je sturmovské právě tehdy, když každý jeho faktor má právě dvě návratová slova. Z toho ihned plyne, že derivovaná slova sturmovských slov jsou opět sturmovská.

Sturmovská slova mají řadu zobecnění na vícepísmenné abecedy, jedním z nich jsou i Arnoux–Rauzyho slova. Uniformně rekurentní nekonečné slovo $\mathbf{u}$ nad abecedou $\mathcal{A}$ je *Arnoux–Rauzyho*, pokud má pro každé n právě $(\#\mathcal{A} - 1)n + 1$ faktorů délky n a pro každou délku n má právě jeden levý a jeden pravý speciální faktor. Autoři práce [2] ukázali, že každý faktor Arnoux–Rauzyho slova nad abecedou $\mathcal{A}$ má právě $\#\mathcal{A}$ návratových slov. To znamená, že derivovaná slova Arnoux–Rauzyho slov nad $\mathcal{A}$ mohou být uvažována nad stejnou abecedou $\mathcal{A}$. Na rozdíl od sturmovských slov ale tato vlastnost necharakterizuje Arnoux–Rauzyho slova, pokud $\#\mathcal{A} > 2$.

Cílem tohoto článku je studovat derivovaná slova Arnoux–Rauzyho slov. Pro každé Arnoux–Rauzyho slovo $\mathbf{u}$ popíšeme množinu $\mathrm{Der}(\mathbf{u})$ všech derivovaných slov k neprázdným prefixům slova $\mathbf{u}$. Přímým důsledkem této charakterizace je fakt, že každé derivované slovo Arnoux–Rauzyho slova je opět Arnoux–Rauzyho slovo. Durand [1] ukázal, že posloupnost $\mathbf{u}$ je primitivně substitutivní, tj. je to morfický obraz pevného bodu primitivního morfismu, právě tehdy, když množina $\mathrm{Der}(\mathbf{u})$ je konečná. V tomto článku stanovujeme přesnou velikost množiny $\mathrm{Der}(\mathbf{u})$ pro všechny primitivně substitutivní Arnoux–Rauzyho slova. Tím zobecňujeme předchozí výsledky [3], které odhadují velikost množiny $\mathrm{Der}(\mathbf{u})$ pro pevné body primitivních sturmovských morfismů.

Klíčová slova: Arnoux–Rauzyho slovo, derivované slovo, návratové slovo

Full paper: K. Medková. *Derived sequences of Arnoux-Rauzy sequences.* In 'Combinatorics on Words (WORDS, Loughborough, 2019)', R. Mercaş and D. Reidenbach (eds.), volume 11682 of *Lecture Notes in Computer Science*, Springer (2019), 251–263.

References

[1] F. Durand. *A characterization of substitutive sequences using return words.* Discrete Math. **179** (1998), 89–101.

[2] J. Justin and L. Vuillon. *Return words in Sturmian and episturmian words.* RAIRO-Theoret. Inf. Appl. **34** (2000), 343–356.

[3] K. Klouda, K. Medková, E. Pelantová and Š. Starosta. *Fixed points of Sturmian morphisms and their derivated words.* Theoret. Comput. Sci. **743** (2018), 23–37.

[4] L. Vuillon. *A characterization of Sturmian words by return words.* Eur. J. Combin. **22** (2001), 263–275.

On Long-Term Properties of Geometric Flows of Space Curves*

Jiří Minarčík

2nd year of PGS, email: minarji2@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. This contribution deals with problems associated with generalization of the curve shortening flow into higher dimensional space. Specifically, the motion in normal and binormal direction of closed curves embedded in $\mathbb{R}^3$ is analyzed and compared to the standard two-dimensional case. Although the motion of space curves in normal direction is similar to that in the plane, new phenomena may be observed in $\mathbb{R}^3$. Namely, embedded curves may develop new types singularities, stop being simple or loose their convexity during the evolution. We discuss some specific examples and present theoretical results addressing these problems. The motion in the binormal direction, discussed in the second part of the talk, has been studied mainly in the context of vortex dynamics. Aside from its application in physics, this motion has many interesting properties. The local rate of parametrization as well as length, total torsion, elastic energy and other global properties of closed curves are preserved and the motion law is equivalent to a non-linear Schrödinger equation obtained by the Hasimoto's transformation.

Keywords: space curves, geometric flow, parametric approach

Abstrakt. Tento příspěvek poukazuje na problémy spojené se zobecněním geometrických toků křivek. Zkoumán je především pohyb uzavřených prostorových křivek, který je srovnán s klasickým pohybem v rovině. Narozdíl od běžné definice, může pohyb jednoduchých křivek v $\mathbb{R}^3$ generovat komplikovanější typy singularit, části křivky do sebe mohou narazit a z původně konvexních křivek se mohou stát křivky konkávní. Tyto problémy jsou nejprve teoreticky analyzovány a poté vysvětleny na konkrétních příkladech. Uvažován je také pohyb ve směru binormálního vektoru, který vede k jednoduchému modelu dynamiky vírových smyček v Eulerovské kapalině.

Klíčová slova: prostorové křivky, geometrický tok, parametrická formulace

References

[1] J. Minarčík, M. Kimura, M. Beneš. *Comparing Motion of Curves and Hypersurfaces in* $\mathbb{R}^m$. Discrete and Continuous Dynamical Systems Series B, 2019 (24), 4815–4826.

[2] J. Minarčík, M. Beneš. *Long-term properties of curve shortening flow in* $\mathbb{R}^3$. Submitted to SIAM Journal on Mathematical Analysis, 2019.

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/194/OHK4/3T/14

Knowledge-Based Selection of Gaussian Process Surrogates in Combination with the CMA-ES*

Zbyněk Pitra†

5th 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 Machine Learning
Institute of Computer Science, CAS

Abstract. Many real-world problems belong to the area of continuous black-box optimization. If the black-box function is also expensive, regression surrogate models are often utilized by optimization algorithms to save evaluations of the original expensive function. In such optimization tasks, Gaussian processes [7] modeling technique has been shown as a valuable surrogate model for the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [3], the state-of-the-art algorithm in continuous black-box optimization field. Choosing a suitable surrogate model or a suitable setting of its hyperparameters is a complex selection problem, where research into reusing knowledge represented by features of black-box function landscape is only starting.

In [6], we investigated how different Gaussian process settings influence the error between the predicted and genuine population ordering in connection with features representing the fitness landscape. Apart from using features for landscape analysis known from the literature, we proposed a new set of features based on CMA-ES state variables. We performed the landscape analysis of a large set of data generated using runs of a surrogate-assisted version of the CMA-ES, the Doubly Trained Surrogate CMA-ES [1, 4] (DTS-CMA-ES), on the noiseless part of the Comparing Continuous Optimisers [2] benchmark function testbed.

The promising results of the previous research lead to the investigation into surrogate model selection [5], where we utilized the knowledge from the previous model experience to design a metalearing system. As a proof of concept, we provided a study investigating the influence of landscape features on the performance of various Gaussian process covariance functions as surrogate models for the DTS-CMA-ES.

Keywords: black-box optimization, surrogate model, Gaussian process, metalearing

Abstrakt. Mnoho praktických úloh spadá do oblasti spojité black-box optimalizace. Pokud je black-box funkce navíc drahá, optimalizační algoritmy využívají namísto původní funkce náhradní regresní model. Gaussovské procesy prokázaly, že jsou vhodným náhradním modelem pro algoritmus Covariance Matrix Adaptation Evolution Strategy [3] (CMA-ES) při řešení optimalizačních úlohy tohoto typu. Výběr vhodného náhradního modelu nebo vhodného nastavení jeho parametrů je komplexní problém výběru, jehož výzkum v oblasti opakovaného využití

*The reported research was supported by the Czech Science Foundation grants Nos. 17-01251S and 18-18080S and by the Grant Agency of the Czech Technical University in Prague with its grant No. SGS17/193/OHK4/3T/14.

†This study has been provided in cooperation with Lukáš Bajer and Jakub Repický.

předchozích znalostí, jež jsou reprezentovány příznaky tvaru black-box funkce, je teprve na svém počátku.

Ve článku [6] jsme zkoumali, jak různá nastavení gaussovských procesů ovlivňují chybu předpovězeného pořadí populace vůči pořadí skutečnému ve spojení s příznaky představujícími tvar fitness funkce. Kromě příznaků známých z literatury jsme také představili zcela novou množinu příznaků založenou na stavových proměných algoritmu CMA-ES. Provedli jsme analýzu tvaru funkcí na velkém množství dat vytvořených za použití běhů verze algoritmu CMA-ES využívající náhradní model, jmenovitě algoritmu Doubly Trained Surrogate CMA-ES [1, 4] (DTS-CMA-ES), na funkcích z platformy Comparing Continuous Optimisers [2] bez přidaného šumu.

Slibné výsledky předchozího výzkumu nás přivedly ke zkoumání výběru náhradního modelu [5], kde jsme využívali předchozích zkušeností modelu pro vytvoření metaučícího systému. Abychom ověřili náš koncept, provedli jsme studii vlivu příznaků tvaru funkce na úspěšnost predikce různých kovariančních funkcí gaussovského procesu jakožto náhradního modelu pro DTS-CMA-ES.

Klíčová slova: black-box optimalizace, náhradní modelování, gaussovské procesy, metaučení

References

[1] L. Bajer, Z. Pitra, J. Repický, and M. Holeňa. *Gaussian process surrogate models for the CMA Evolution Strategy.* Evolutionary Computation **0** (0), 1–33. PMID: 30540493.

[2] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Technical Report RR-6829, INRIA, (2009). Updated February 2010.

[3] 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.

[4] Z. Pitra, L. Bajer, and M. Holeňa. *Doubly Trained Evolution Control for the Surrogate CMA-ES*, 59–68. Springer International Publishing, Cham, (2016).

[5] Z. Pitra, L. Bajer, and M. Holeňa. Knowledge-based selection of Gaussian process surrogates. In 'ECML PKDD 2019: Workshop on Interactive Adaptive Learning. Proceedings', D. Kottke, V. Lemaire, A. Calma, G. Krempl, and A. Holzinger, (eds.), ECML PKDD 2019, 48–63, Würzburg, Germany, (September 2019).

[6] Z. Pitra, J. Repický, and M. Holeňa. Landscape analysis of Gaussian process surrogates for the Covariance Matrix Adaptation Evolution Strategy. In 'Proceedings of the Genetic and Evolutionary Computation Conference', GECCO '19, 691–699, New York, NY, USA, (2019). ACM.

[7] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* Adaptative computation and machine learning series. MIT Press, (2006).

Transcendental Transport System Control

Peter Pribeli

1st year of PGS, email: pribepe1@fjfi.cvut.cz.cz
Department of Mathematics
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. Dynamical systems appear in a plethora of applications where the linear part possesses a transcendental transport function in the Laplace, Fourier or Z-transform. This is caused by the form of the linear model which usually entails non-constant coefficients, damping terms or is modelled by partial differential equations of the hyperbolic or parabolic type. The presence of non-integer orders of derivatives in the model of the anomalous system behaviour further complicates matters. The thesis' objective is to explore these models by investigating model formulation, continuous and discrete transport and the design of a control algorithm. The goal is to study the transport functions by means of mathematical analysis in the complex domain, find an analytical solution or its approximation and simulate the systems in order to verify their properties.

Keywords: Transcendental transport, Laplace transform, Fourier transform

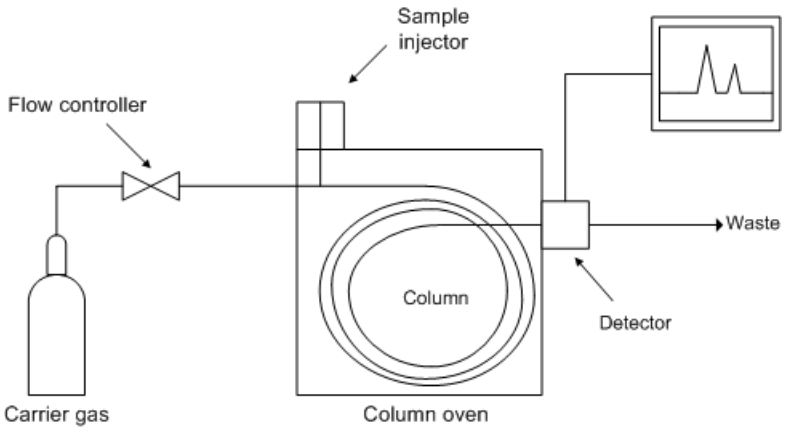
Abstrakt. V celé řadě aplikací se vyskytují dynamické systémy, jejichž lineární část má transcendentní přenosovou funkci v Laplaceově, Fourierově či Z-transformaci. Příčinou je tvar lineárního modelu, který obvykle obsahuje nekonstantní koeficienty, zpožďovací členy nebo je popsán parciálními diferenciálními rovnicemi hyperbolického či parabolického typu. Další komplikací je výskyt neceločíselných derivací při popisu anomálního chování systémů. Cílem práce je studium takových systémů od formulace modelu, přes vlastnosti spojitého přenosu až k diskrétnímu přenosu a návrhu řídicího algoritmu. Cílem je studovat přenosové funkce nástroji matematické analýzy v komplexním oboru, nalézt analytické řešení nebo jeho aproximaci a provézt simulační výpočty pro ověření vlastností systémů.

Klíčová slova: Transcendentální přenos, Laplaceova transformace, Fourierova transformace

1 Introduction

This contribution explores the numerical inversion of the laplace transform. The motivation of the investigation of the numerical inversion of the Laplace transform is laid out below.

Consider a gas chromatograph. This device, depicted in Figure 1, employs an elongated column to separate gas mixtures. The column is layered with diffusive compounds in order to aid separation.



A depiction of the gas chromatograph[1].

2 General model

Suppose there are $n \in \mathcal{N}$ components in the gas phase on the inlet and $N \in \mathcal{N}$ in the diffusive phase. Assuming $1st$ order kinematics, the batch kinematics of the system obey

$$\frac{\partial \vec{c}}{\partial t} = \mathbb{P}\vec{c} + \mathbb{Q}\vec{a}, \tag{1}$$

$$\frac{\partial \vec{a}}{\partial t} = \mathbb{R}\vec{c} + \mathbb{S}\vec{a}, \tag{2}$$

where the matrices $\mathbb{P}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{S}$ describe the batch kinematics of the system and $c_k(x,t),\ k \in 1,\dots,n$ and $a_k(x,t),\ k \in 1,\dots,N$ are the concentrations in the gas phase and in the stationary phase, respectively. The so-called *piston model* introduces a term in the equation which models the gas phase as a piston propagating through the column, are governed by the system of equations

$$\frac{\partial \vec{c}}{\partial t} = -v\frac{\partial \vec{c}}{\partial x} + \mathbb{P}\vec{c} + \mathbb{Q}\vec{a}, \tag{3}$$

$$\frac{\partial \vec{a}}{\partial t} = \mathbb{R}\vec{c} + \mathbb{S}\vec{a}, \tag{4}$$

with constrains and initial conditions

$$t > 0,\ 0 < x < L, \tag{5}$$

$$\vec{c}(0^+,t) =: \vec{u}(t), \tag{6}$$

$$\vec{c}(L^-,t) =: \vec{y}(t), \tag{7}$$

$$\vec{c}(x,0^+) = \vec{0}, \tag{8}$$

$$\vec{a}(x,0^+) = \vec{0}, \tag{9}$$

where L is the length of the column and v is the gas velocity.

In order to solve the aforementioned problem, the Laplace transform

$$\vec{C}(x,s) = \mathscr{L}\{\vec{c}(x,t)\}, \tag{10}$$

$$\vec{A}(x,s) = \mathscr{L}\{\vec{a}(x,t)\} \tag{11}$$

is applied to both sides of equations (3) and (4) yielding

$$s\vec{C} = -v\frac{\partial \vec{C}}{\partial x} + \mathbb{P}\vec{C} + \mathbb{Q}\vec{A} \;\Rightarrow\; \frac{\partial \vec{C}}{\partial x} = \frac{1}{v}\left(\mathbb{P} + \mathbb{Q}\left(\mathbb{I}s - \mathbb{S}\right)^{-1}\mathbb{R}\right)\vec{C}, \tag{12}$$

$$s\vec{A} = \mathbb{R}\vec{c} + \mathbb{S}\vec{a} \;\Rightarrow\; \vec{A} = \left(\mathbb{I}s - \mathbb{S}\right)^{-1}\mathbb{R}\vec{C}, \tag{13}$$

with boundary conditions

$$\vec{C}(0^+, t) =: \vec{U}(t), \tag{14}$$

$$\vec{C}(L^-, t) =: \vec{Y}(t), \tag{15}$$

$$\tag{16}$$

Hence, the general solution to (12) and (13) is

$$\vec{C}(x,t) = \exp\left[\frac{x}{v}\left(\mathbb{P} + \mathbb{Q}\left(\mathbb{I}s - \mathbb{S}\right)^{-1}\mathbb{R} - \mathbb{I}s\right)\right]\vec{c}(x,t), \tag{17}$$

$$\vec{Y}(x,t) = \exp\left[\underbrace{\frac{L}{v}}_{T}\left(\mathbb{P} + \mathbb{Q}\left(\mathbb{I}s - \mathbb{S}\right)^{-1}\mathbb{R} - \mathbb{I}s\right)\right]\vec{c}(x,t). \tag{18}$$

One can see that the homogeneous solution is

$$\mathbb{F}(s) = \exp\left\{\left(\mathbb{P} + \mathbb{Q}\left(\mathbb{I}s - \mathbb{S}\right)^{-1}\mathbb{R} - \mathbb{I}s\right)\right\} = \exp(-Ts)\exp(\mathbb{P}T)\exp\left(\mathbb{Q}\left(\mathbb{I}s - \mathbb{S}\right)^{-1}\mathbb{R}T\right). \tag{19}$$

3 One equilibrium reaction

In order to clarify the above model, consider an equilibrium equation with two compounds with concentrations c_1 and c_2

$$c_1 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} a_1. \tag{20}$$

The system of equations (1) and (2) are now transformed to

$$\frac{\mathrm{d}c_1}{\mathrm{d}t} = -k_1 c_1 + k_2 a_1, \tag{21}$$

$$\frac{\mathrm{d}a_1}{\mathrm{d}t} = k_1 c_1 - k_2 a_1, \tag{22}$$

Since the equations no longer involve matrices, one can readily express the fundamental solution as

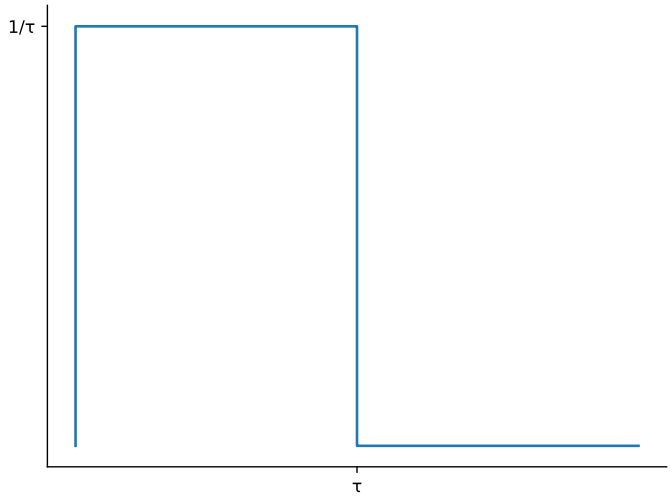
$$F(s) = \exp\left(-Ts\right)\exp\left(-k_1 T\right)\exp\left(k_2\left(s + k_2\right)^{-1}k_1 T\right), \tag{23}$$

the inverse matrices reduce to an inversion of a scalar

$$F(s) = \exp\left(-Ts\right)\exp\left(-k_1 T\right)\exp\left(\frac{k_1 k_2 T}{s + k_2}\right). \tag{24}$$

Observe that the solution has an essential singularity at $s = -k_2$. This singularity would correspond to an eigenvalue of the matrix $\mathbb{S}$.

Note that the foregoing construction absents a piston term $-v\frac{\partial c_1}{\partial x}$. This is due to the fact that a Dirac inlet pulse is assumed and since the Laplace transform of a Dirac delta function is constant, the derivative is omitted.



The step function used as a boundary condition for the inlet of a chromatograph.

3.1 Realistic inlet

The Dirac delta function does not represent a realistic inlet. Therefore a more realistic inlet in the form of a normalized pulse depicted in Figure 3.1 is considered. The normalized stepfunction is introduced by the boundary condition

$$u(t) = \frac{1 - I(t \geq \tau)}{\tau}, \tag{25}$$

where I is the unit function. The Laplace transform of the boundary condition is

$$U(s) = \frac{1 - \exp(-\tau s)}{\tau s}. \tag{26}$$

Employing (18) the solution can be expressed as

$$Y(t) = \frac{\exp\left(-k_1 T\right) \exp\left(-Ts\right)}{\tau} \left(1 - \exp\left(-\tau\right)\right) \exp\left(\frac{k_1 k_2 T}{s + k_2}\right). \tag{27}$$

In order to obtain a solution of the outlet, one has to invert the Laplace transform in (27). The relation can bu further simplified as

$$y(t) = \frac{\exp\left(-k_1 T\right)}{\tau} \left(\varphi\left(t - T\right) - \varphi\left(t - T - \tau\right)\right), \tag{28}$$

where a new function φ has been introduced in the form

$$\varphi(t) = \begin{cases} 0, & t \leq 0, \\ \mathscr{L}^{-1}\left\{\frac{1}{s} \exp\left(\frac{k_1 k_2 T}{s + k_2}\right)\right\}, & t > 0. \end{cases} \tag{29}$$

Hence the problem involves computing an inverse laplace transformation of (29).

4 Inverse Laplace Transform

On can show that an inverse Laplace transform is equivalent to computing the Bromwich integral

$$\mathscr{L}^{-1}F(s) = f(t) = \frac{1}{2\pi i}\lim_{T\to\infty}\int_{\gamma-iT}^{\gamma+iT}\exp(st)F(s)ds, \tag{30}$$

where $t > 0$. The contour along which the above integral is calculated is chosen such that the real part of all the singularities of $F(s)$ lie to the left of the contour. One can devise various approaches to numerically calculate the inverse Laplace transform.

4.1 Talbot

The first method examined is due to Talbot[2, 3]. Talbot's approach utilises the fact that the Bromwich contour in (30) can be arbitrarily deformed due to Cauchy's theorem as long as the new contour is analytical. In order to improve the convergence of (30), the integral needs values of s with a large negative real component. The fixed Talbot algorithm changes the contour of integration to

$$s(\theta) = \gamma\theta(\cot\theta + i),\ -\pi < \theta < \pi, \tag{31}$$

where γ is a parameter. Note that the parameter has to be chosen in such a way as to not intercept any poles of $F(s)$. The Talbot inverse Laplace therefore takes the form

$$f(t) = \frac{1}{2\pi i}\int_{-\pi}^{\pi}\exp\left(ts(\theta)\right)F(s(\theta))s'(\theta)\mathrm{d}\theta, \tag{32}$$

where $s'(\theta) = i\gamma(1 + i\sigma(\theta))$ and $\sigma(\theta) = \theta + (\theta\cot\theta - 1)\cot\theta$. Evaluating the integral simplifies it to

$$f(t) = \frac{\gamma}{\pi}\int_{0}^{\pi}\mathrm{Re}\left[\exp\left(ts(\theta)\right)F(\theta)(1 + i\sigma(\theta))\right]\mathrm{d}\theta. \tag{33}$$

This integral is calculated numerically with the aid of a trapezoidal approximation with step size π/M and $\theta_k = k\pi/M$

$$f(t, M) = \frac{\gamma}{M}\left\{\frac{1}{2}F(\gamma)\exp(\gamma t) + \sum_{k=1}^{M-1}\mathrm{Re}\left[\exp\left(ts(\theta_k)\right)F\left(s(\theta_k)\right)\left(1 + i\sigma\left(\theta_k\right)\right)\right]\right\}. \tag{34}$$

The value of γ has been determined empirically[3] to be

$$\gamma = \frac{2M}{5t}. \tag{35}$$

The algorithm is called the fixed Talbot algorithm due to the fact that the contour is fixed by the value of γ. The approximation (34) has only one free parameter, M, which is the number of terms to be summed up. Ergo it represents the precision of the approximation.

4.2 Gaver-Stehfest

Gaver proposed and prooved[4] a successive approximation of $f(t)$ with the set of functions

$$f_k(t) = \int_0^\infty p_k(s) f\left(\frac{ts}{\ln 2}\right) \mathrm{d}s, \tag{36}$$

where

$$p_k(s) = \frac{(2k)!}{k!(k-1)!}(1-\exp(-s))^k \exp(-ks),\ k \geq 1,\ s \geq 0. \tag{37}$$

Gaver proved the convergence

$$\lim_{k\to\infty} f_k(t) = f(t), \tag{38}$$

however realized that the convergence is very slow. Stehfest used[5, 6], a convergence acceleration technique to accelerate the convergence of (38). He introduced new asymptotic coefficients and proposed an approximation known as the Gaver-Stehfest method

$$f_n(t) = \frac{\ln 2}{t} \sum_{k=1}^{2n} a_k(n) F\left(\frac{k \ln 2}{t}\right), \tag{39}$$

where

$$a_k(n) = \frac{(-1)^(n+k)}{n!} \sum_{j=[(k+1)/2]}^{\min(k,N/2)} j^{n+1} \binom{n}{j} \binom{2j}{j} \binom{j}{k-j}, \tag{40}$$

where $n \geq 1$ and $1 \leq k \leq 2n$.

4.3 de Hoog-Knight-Stokes

This algorithm is based on the fact that the inverse (30) can be decomposed into a Fourier series. By considering the real and imaginary parts separately one arrives at

$$f(t) = \frac{2}{\pi} \exp(\gamma t) \int_0^\infty \mathrm{Re}\left[F(s)\right] \cos(\omega t) \mathrm{d}\omega \tag{41}$$

$$= -\frac{2}{\pi} \exp(\gamma t) \int_0^\infty \mathrm{Im}\left[F(s)\right] \sin(\omega t) \mathrm{d}\omega \tag{42}$$

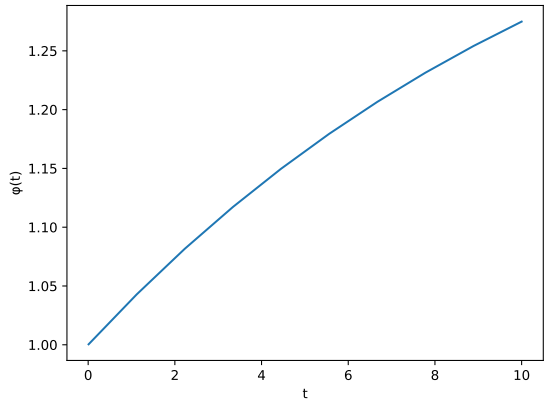
$$= \frac{1}{\pi} \exp(\gamma t) \int_0^\infty \mathrm{Re}\left[F(s)\right] \exp(i\omega t) \mathrm{d}\omega. \tag{43}$$

The third equation is usually chosen for the numerical approximations. Using the trapezoidal rule with step of size π/T the Fourier integral is approximated as

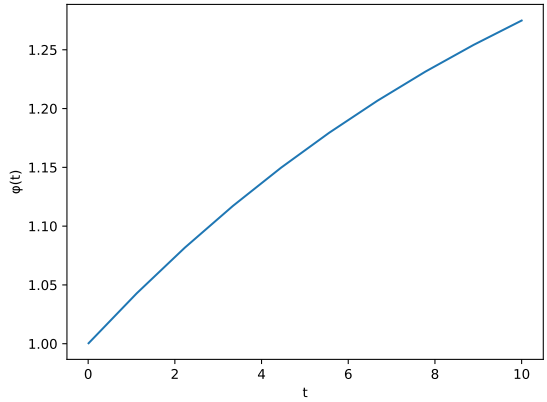
$$f(t, 2M+1) = \frac{\exp(\gamma t)}{T} \sum_{k=0}^{2M} \mathrm{Re}\left[F(s_k) \exp\left(\frac{i\pi t}{T}\right)\right]. \tag{44}$$

This sum however converges very slowly. Employing a Pade acceleration schema, the authors propose[7] an accelerated estimate

$$f(t, \gamma, T, M) = \frac{1}{T} \exp(\gamma t) \mathrm{Re}\left[\varepsilon_{2M}^0\right], \tag{45}$$



The Talbot Inverse Laplace transform.



The Gaver-Stehfest Inverse Laplace transform.

where

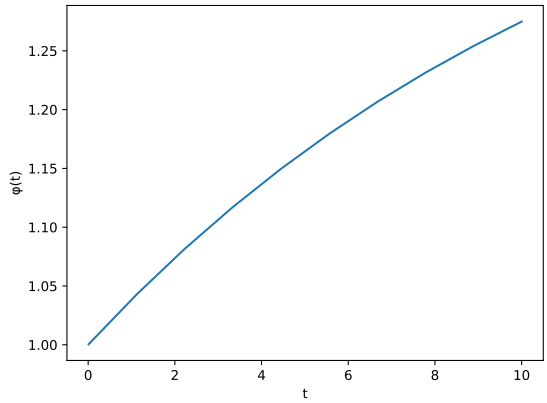
$$\varepsilon_{2M}^{0} = \left[\sum_{n=0}^{m} b_n z^n\right] \Big/ \left[\sum_{n=0}^{m} b_n z^n\right], \; c_0 = 0 \tag{46}$$

are the diagonal elements of the Pade table.

5 Numerical experiments

Numerical experiments of the inversion of (29) have been carried out. All three of the aforementioned algorithms for the inversion of the Laplace transform are implemented in the Python package *mpmath* package[8].

For illustration purposes, the values of the equilibrium constants have been chosen as $k_1 = 0.2$ and $k_2 = 0.1$ and $T = 2$. The preliminary results are shown in figures 55 and 5. No significant difference between the three algorithms has been found since the numerical results are identical. Note that these results are very preliminary and a further tuning shall be carried out.



The de Hoog Inverse Laplace transform.

References

[1] Wikipedia user *Rune.welsh*. Published under the Creative Commons license CC BY-SA 3.0.

[2] A. Talbot *The accurate numerical inversion of Laplace transforms.* IMA Journal of Applied Mathematics 23(1):97, (1979)

[3] J. Abate, P. Valko *Multi-precision Laplace transform inversion.* International Journal for Numerical Methods in Engineering 60:979-993 (2004)

[4] D. Gaver, Jr. *Observing stochastic processes, and approximate transform inversion.* Operations Research, 14(3):pp. 444–459, (1966)

[5] H. Stehfest. *Algorithm 368: Numerical inversion of Laplace transforms.* Commun. ACM, 13(1):47– 49, (1970)

[6] H. Stehfest. *Remark on algorithm 368: Numerical inversion of laplace transforms.* Commun. ACM, 13(10):624, (1970)

[7] F. de Hoog, J. Knight, A. Stokes *An improved method for numerical inversion of Laplace transforms.* SIAM Journal of Scientific and Statistical Computing 3:357-366 (1982)

[8] F. Johansson *et al mpmath: a Python library for arbitrary-precision floating-point arithmetic* (version 0.18) (2019)

Borland's Process in Incomplete Market*

Martin Prokš

3rd year of PGS, email: proksma6@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. Possible generalizations of the Borland process are discussed in this paper. In the original Borland model, constant volatility was assumed which leads to a complete market. This restriction, however, is not supported by empirical observations. Therefore, it is attempted to relax this assumption here and hence approach a more realistic scenario occurring in a financial market. Introducing another source of randomness brings various complications, mainly transition from complete to incomplete market. All those difficulties are addressed here. The advantage of an option price obtainable from the generalized model is that it should better compensate a trader for entering into risky contract since a more realistic underlying model is assumed.

Keywords: stochastic process, stochastic volatility, Tsallis distribution, martingale, option pricing

Abstrakt. V této práci jsou diskutovaná možná zobecnění Borland procesu. V originálním Borlandově modelu byla předpokládána konstantní volatilita, tento přepoklad vede na "úplný" trh. Nicméně toto omezení není v souladu s empirickým pozorováním. Proto se v této práci pokusíme tento předpoklad odstranit a tím se přiblížit k realnějšímu případu vyskytujícího se na finančním trhu. Přidání dalšího zdroje náhodnosti přináší mnohé komplikace, hlavně přechod z "úplného" trhu na "neúplný". Všechny tyto obtíže jsou diskutovány. Výhoda ceny opce získatelné ze zobecněného modelu je, že tato cena by měla lépe kompenzovat obchodníka za vstup do riskantního kontraktu, protože je použit reálnější model pro podkladové aktivum.

Klíčová slova: stochastický proces, stachastická volatilita, Tsallisova distribuce, martingale, oceňování opcí

1 Introduction

An option pricing plays an important role in a modern financial industry. The fundamental assumption for finding a correct option price is choosing the most realistic model for an underlying asset. Until recently, the most commonly used model was Black-Scholes, see [9] or any textbook on mathematical finance. However, there is an increasing demand for generalized models going beyond this standard paradigm, see for instance [1]. Lisa Borland proposed one such generalization in her paper [3]. The major drawback of her model, however, is an assumption of constant volatility which is not supported by

*This work has been 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.

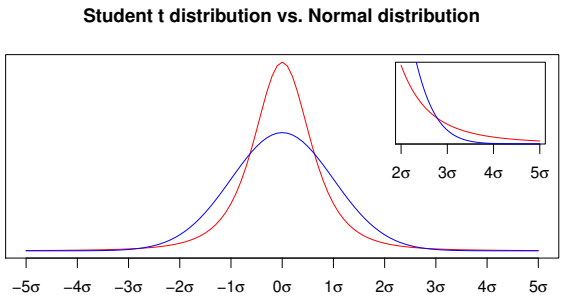


Figure 1: Comparison of Student and Gaussian distributions

empirical observations. In the present paper, the Borland process is recalled and its generalization in form of introducing stochastic volatility is proposed in order to approach a more realistic scenario occurring in financial markets.

Next, the generalized model is applied to find an option price by using a no-arbitrage paradigm. This paradigm leads to a partial differential equation and its solution determines the option price. The advantage of the option price obtainable from the generalized model is that it should better compensate a trader for entering into a risky contract since a more realistic underlying model is assumed.

The structure of this paper is as follows. In section 2, the new model with stochastic volatility is introduced. Next, the standard procedure for finding option pricing formula for this particular model is presented in section 3 together with discussion on incomplete market difficulties. Finally, section 4 suggests ideas for reformulating Borland's model into a realm of semmimartingales which may facilitate derivative pricing under generalized model.

2 Model specification

In this central section, the proposed model is introduced along with a brief discussion of the original model. The original Stochastic model for an evolution of an underlying asset S is

$$\mathrm{d}S_t = \mu S \mathrm{d}t + \sqrt{\nu} S \mathrm{d}\Omega_t, \tag{1}$$

where $\mathrm{d}\Omega_t$ is defined by a stochastic differential equation

$$\mathrm{d}\Omega_t = P_q(\Omega_t)^{\frac{1-q}{2}} \mathrm{d}B_t^{(1)}, \tag{2}$$

and $P_q(\Omega_t)$ is the solution of the Fokker-Planck equation corresponding to equation 2, i.e. the transition probability density,

$$\frac{\partial P_q}{\partial t}(\Omega, t; \Omega', t') = \frac{1}{2} \frac{\partial^2 P_q^{2-q}}{\partial \Omega^2}(\Omega, t; \Omega', t'), \tag{3}$$

$$P_q(\Omega, t; \Omega', t') \to \delta(\Omega - \Omega') \quad as \quad t \to t'. \tag{4}$$

The model specified by equations 1 and 2 is adopted from the original paper [3]. However, this form of writing of the stochastic differential equation in 1 is not strictly

speaking correct since $\mathrm{d}\Omega_t$ is not Brownian noise. A correct description of the model is given by two coupled SDE

$$\mathrm{d}S_t = \mu S \mathrm{d}t + \sqrt{\nu} S P_q(\Omega_t)^{\frac{1-q}{2}} \mathrm{d}B_t^{(1)}, \tag{5}$$

$$\mathrm{d}\Omega_t = P_q(\Omega_t)^{\frac{1-q}{2}} \mathrm{d}B_t^{(1)}. \tag{6}$$

Despite the mathematical incorrectness, there is a reason for writing the model in the original form since it highlights the fact that the non-Brownian model for underlying asset is used. The motivation for introducing non-Brownian noise $\mathrm{d}\Omega_t$ comes from the empirical observation which shows that so-called Tsallis distribution fits empirical log-returns much better than standard normal distribution, see [3]. The justification of the model then comes from the fact that Tsallis distribution is the solution of equation 3 and hence it is the probability distribution of Ω_t (macroscopic behaviour). In addition, it can be shown that $\mathrm{d}\Omega_t$ corresponds to a random part of log-returns. Therefore, this model effectively captures observational behaviour.

Another interesting property is that the model may effectively capture periods of high and low volatility. Imagine driving process Ω_t reaches regions with low probability, then $P_q(\Omega_t)$ becomes very small, and for $q \approx 1.5$, which is observed in a real market, the volatility term turns into $\frac{1}{P_q(\Omega)^{1/4}}$ which becomes larger than in the case of highly probable value of Ω_t. The process Ω_t, and hence also process S_t, will undergo a period of time with higher volatility until the high volatility drives the process Ω_t into more probable regions.

Unlike in the original paper, the volatility ν in the model is considered here to be stochastic, and its dynamics is given by the stochastic differential equation

$$\mathrm{d}\nu_t = a(S, \nu, t)\mathrm{d}t + b(S, \nu, t)\mathrm{d}B_t^{(2)}. \tag{7}$$

The volatility model may for now stay in a full generality, i.e. a and b are arbitrary functions (the only condition is that the equation 7 must have a solution).

The whole generalized model can be written in a compact matrix form

$$\mathrm{d}\mathbf{X} = \begin{pmatrix} \mathrm{d}S \\ \mathrm{d}\Omega \\ \mathrm{d}\nu \end{pmatrix} = \begin{pmatrix} \mu S \\ 0 \\ a(S,\nu,t) \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \sqrt{\nu} S P_q(\Omega)^{\frac{1-q}{2}} & 0 \\ P_q(\Omega)^{\frac{1-q}{2}} & 0 \\ 0 & b(S,\nu,t) \end{pmatrix} \begin{pmatrix} \mathrm{d}B_1 \\ \mathrm{d}B_2 \end{pmatrix}. \tag{8}$$

Note that Itô lemma may still be applied on individual components of the vector stochastic process $\mathbf{X}$. In other words, it is possible to get a simpler equation if S is transformed into $Y_t = f(S, \nu, t)$, where we choose $f(S, \nu, t) = \ln S$. Then, instead of the first equation, we have equation for Y_t

$$\mathrm{d}Y_t = \tilde{\mu}\mathrm{d}t + \sqrt{\nu}\mathrm{d}\Omega_t, \tag{9}$$

where

$$\tilde{\mu} = \mu - \frac{1}{2}\nu P_q^{1-q}(\Omega). \tag{10}$$

3 Option pricing

The main goal in this section is to derive a price for an option C on the underlying asset S. The option formula is derived using a no-arbitrage paradigm. This paradigm dictates to construct a portfolio containing the option and some amount Δ_1 of the underlying asset S. However, it is known from the standard theory that when another source of randomness is present, such as random volatility in our case, some amount of another option Δ_2 must be included into the portfolio Π

$$\Pi = C - \Delta_1 S - \Delta_2 C_2. \tag{11}$$

Subsequently, the option price is obtained by requiring the portfolio to be risk-free. That is its increment must not contain any random term. The increment of the portfolio is given by

$$\mathrm{d}\Pi = \mathrm{d}C - \Delta_1 \mathrm{d}S - \Delta_2 \mathrm{d}C_2, \tag{12}$$

where the increment of the option is obtained from multi-dimensional Itô lemma

$$\begin{aligned}\mathrm{d}C(S,t,\nu) &= \frac{\partial C}{\partial t}\mathrm{d}t + \frac{\partial C}{\partial S}\mathrm{d}S + \frac{\partial C}{\partial \nu}\mathrm{d}\nu + \frac{1}{2}\frac{\partial^2 C}{\partial S^2}\mathrm{d}S^2 + \frac{1}{2}\frac{\partial^2 C}{\partial \nu^2}\mathrm{d}\nu^2 + \frac{\partial^2 C}{\partial S \partial \nu}\mathrm{d}S\mathrm{d}\nu \\ &= \left(\frac{\partial C}{\partial t} + \frac{1}{2}\nu S^2 P_q(\Omega)^{1-q}\frac{\partial^2 C}{\partial S^2} + \frac{1}{2}b^2\frac{\partial^2 C}{\partial \nu^2} + \sqrt{\nu}SbP_q(\Omega)^{\frac{1-q}{2}}\rho\frac{\partial^2 C}{\partial S\partial \nu}\right)\mathrm{d}t + \frac{\partial C}{\partial S}\mathrm{d}S + \frac{\partial C}{\partial \nu}\mathrm{d}\nu.\end{aligned} \tag{13}$$

Plugging 13 into 12 gives the final expression for portfolio increment $\mathrm{d}\Pi$, which can be simplified into

$$\mathrm{d}\Pi = A\mathrm{d}t + \left(\frac{\partial C}{\partial S} - \Delta_1 - \Delta_2\frac{\partial C_2}{\partial S}\right)\mathrm{d}S + \left(\frac{\partial C}{\partial \nu} - \Delta_2\frac{\partial C_2}{\partial \nu}\right)\mathrm{d}\nu, \tag{14}$$

where the deterministic part is given by

$$\begin{aligned}A := \frac{\partial C}{\partial t} + \frac{1}{2}\nu S^2 P_q(\Omega)^{1-q}\frac{\partial^2 C}{\partial S^2} + \frac{1}{2}b^2\frac{\partial^2 C}{\partial \nu^2} + \sqrt{\nu}SbP_q(\Omega)^{\frac{1-q}{2}}\rho\frac{\partial^2 C}{\partial S\partial \nu} + \\ - \Delta_2\left(\frac{\partial C_2}{\partial t} + \frac{1}{2}\nu S^2 P_q(\Omega)^{1-q}\frac{\partial^2 C_2}{\partial S^2} + \frac{1}{2}b^2\frac{\partial^2 C_2}{\partial \nu^2} + \sqrt{\nu}SbP_q(\Omega)^{\frac{1-q}{2}}\rho\frac{\partial^2 C_2}{\partial S\partial \nu}\right)\end{aligned} \tag{15}$$

The next step is to eliminate randomness, i.e. all random terms needs to vanish. This is achieved by requiring the coefficients of $\mathrm{d}S$ and $\mathrm{d}\nu$ in 14 to be equal to zero. This condition gives constrains on Δ_1 and Δ_2.

$$\frac{\partial C}{\partial \nu} - \Delta_2\frac{\partial C_2}{\partial \nu} = 0 \tag{16}$$

and

$$\frac{\partial C}{\partial S} - \Delta_1 - \Delta_2\frac{\partial C_2}{\partial S} = 0. \tag{17}$$

Thus,

$$\Delta_2 = \frac{\frac{\partial C}{\partial \nu}}{\frac{\partial C_2}{\partial \nu}}, \qquad \Delta_1 = \frac{\partial C}{\partial S} - \Delta_2\frac{\partial C_2}{\partial S} = \frac{\partial C}{\partial S} - \frac{\frac{\partial C}{\partial \nu}}{\frac{\partial C_2}{\partial \nu}}\frac{\partial C_2}{\partial S}. \tag{18}$$

The expressions in 18 serves two purposes. First, it helps to express an increment in the portfolio Π. Second, it gives a trading strategy for hedging against possible loss.

According to no-arbitrage principle any risk-free portfolio must grow with risk-free interest rate r, i.e. it must satisfy the equation

$$\mathrm{d}\Pi = r\Pi \mathrm{d}t \tag{19}$$

which gives the equation

$$A = r\left(C - \Delta_1 S - \Delta_2 C_2\right), \tag{20}$$

where A is an abbreviation for 15.

Unlike in the standard Black-Scholes model, this equation does not provide a way for finding the price for the option C. It is just a relation between two unknown functions C and C_2. It can, however, be rearranged into

$$\frac{\frac{\partial C}{\partial t} + \frac{1}{2}\nu S^2 P_q(\Omega)^{1-q}\frac{\partial^2 C}{\partial S^2} + \frac{1}{2}b^2\frac{\partial^2 C}{\partial \nu^2} + \sqrt{\nu}SbP_q(\Omega)^{\frac{1-q}{2}}\rho\frac{\partial^2 C}{\partial S \partial \nu} - rC + rS\frac{\partial C}{\partial S}}{\frac{\partial C}{\partial \nu}} =$$
$$\frac{\frac{\partial C_2}{\partial t} + \frac{1}{2}\nu S^2 P_q(\Omega)^{1-q}\frac{\partial^2 C_2}{\partial S^2} + \frac{1}{2}b^2\frac{\partial^2 C_2}{\partial \nu^2} + \sqrt{\nu}SbP_q(\Omega)^{\frac{1-q}{2}}\rho\frac{\partial^2 C_2}{\partial S \partial \nu} - rC_2 + rS\frac{\partial C_2}{\partial S}}{\frac{\partial C_2}{\partial \nu}} \tag{21}$$

where the expressions for Δ_1 and Δ_2, equation 18, were used.

Since the left hand side of the equation 21 depends only on the parameters (e.g. time to maturity and strike) of the original option C and the right side depends only on the parameters of the other option C_2, it can be concluded that both sides are independent of the parameters of the options. Therefore, both sides can be function only of S, t and ν. This insight gives the equation for the option price C

$$\frac{\partial C}{\partial t} + \frac{1}{2}\nu S^2 P_q(\Omega)^{1-q}\frac{\partial^2 C}{\partial S^2} + \frac{1}{2}b^2\frac{\partial^2 C}{\partial \nu^2} + \sqrt{\nu}SbP_q(\Omega)^{\frac{1-q}{2}}\rho\frac{\partial^2 C}{\partial S \partial \nu} - rC + rS\frac{\partial C}{\partial S} = f(S,t,\nu), \tag{22}$$

where f is some function of S, t and ν. It is common to introduce a new function $\lambda = \lambda(S,t,\nu)$ so that

$$f(S,t,\nu) = -\Big(a(S,t,\nu) - \lambda(S,t,\nu)b(S,t,\nu)\Big)\frac{\partial C}{\partial \nu}(S,t,\nu). \tag{23}$$

The new function λ is then called market price of (volatility) risk, see [9] or [2]. This additional degree of freedom is reflected in non-uniqueness of martingale measure and in financial literature is referred to as market incompleteness. It is typically assumed that market has chosen a specific martingale measure which corresponds to option prices and λ is chosen accordingly.

Another method for option pricing is a martingale method. The price is then determined, see e.g. [6], as

$$C(t) = E^{\mathcal{Q}}\left[\frac{\beta(t)}{\beta(T)}Payoff(T)|\mathcal{F}_t\right], \tag{24}$$

where $\frac{\beta(t)}{\beta(T)}$ is often just a discount factor $e^{-(T-t)r}$ and $\mathcal{Q}$ denotes martingale (risk-free) measure. If the stock price model is specified as a SDE then according to Feynman-Kac theorem, see [2], eq. 22 and 24 are equivalent. The advantage of the martingale

pricing method is its generality. In particular it holds even for processes which are not represented by SDE.

Incomplete market remark

The freedom in choice of function f, respective λ, was to be expected since according to *"Meta Theorem"* [2] the number of random sources $R = 2$ is not equal to the number of traded underlying assets $M = 1$ and hence the market is incomplete.

The approach taken above is similar to portfolio replication. The idea is that if a claim can be replicated by a self-financing portfolio then the price of the claim and the value of the portfolio coincide. The problem here is that the portfolio in 11 is not self-financing because it includes not only risk free asset and the underlying stock but also another option. This is in agreement with incompleteness of the market, market is complete iff every claim can be replicated by a self-financing portfolio, see [6] and since the market model is incomplete we cannot construct a replicating portfolio.

However, the trick with adding another option into the portfolio relies on the fact that the martingale measure (i.e. function f or λ) is the same for all traded claims. And if we choose the price of one benchmark derivative then all other derivatives prices are uniquely determined. The benchmark price is found form the market and by inverting option price formula, the martingale measure chosen by the market is obtained.

4 Possible reformulations of Borland's model

As indicated in the previous section, the Borland's model is not formulated very well. This later leads to complications in its generalization. In the following we try to suggest a possible reformulations of Borland model, namely, using Levy process, subordinates or change of measure. All those methods are included in a broad class of semmimartingale models of financial market. That is the resulting stock price process may be written in the form, see [8]

$$X(t) = X(0) + A(t) + M(t), \tag{25}$$

where $A(t)$ is a stochastic process with finite variation and $M(t)$ is a local martingale.

4.1 Levy process formulation

Since a crucial characteristic of Borland's model is its Tsallis distribution for log returns. One may try to achieve a similar model with Levy process. Since Student distribution (Tsallis) is infinitely divisible, a Levy process with given $X(1) \sim \mathit{Student}$ shall preserve the desired feature. After reformulation of Borland process into Levy process, the generalization to stochastic volatility may be feasible since stochastic volatility under Levy model is a text book material, see [4].

Unfortunately, there are two complications. First, Student distribution is not stable. Therefore a time scale at which the increments follow Student distribution needs to be fixed and on any other time scale the distribution is different. Second, its mathematical structure is fairly complicated, see [5].

4.2 Subordination

Another feasible reformulation is using Subordinate process. It is a well known fact that every Levy process can be written as a subordination of a standard Brownian motion, see e.g. [7]. The difficult part is inverting this relation, i.e. finding a proper random time such that we obtain a desired Levy process after subordination. This overlaps with the previous generalization. The only difference and advantage would be staying in a well known regime of Brownian motion.

$$X(t) = B(\tau(t)), \tag{26}$$

where $\tau(t)$ is a non-decreasing Levy process.

4.3 Change of Time

The concept of change of time extends the subordination by relaxing the assumption that the new time τ needs to be Levy process. Moreover, it introduces a stochastic integral representation of subordination, see [7],

$$X(t) = B(\tau(t)) = \int_0^t H_s \mathrm{d}B_s. \tag{27}$$

Hence we can write

$$\mathrm{d}X_t = H_t \mathrm{d}B_t \tag{28}$$

and call process H_t a derivative of new time with respect to the old time.

5 Conclusion

The partial differential equation for the option price was derived under the condition that the underlying asset follows the Borland process with a stochastic volatility. The derivation followed the standard option pricing paradigm, i.e. no-arbitrage argument. The proposed generalized model should better capture the behaviour of stock prices, and therefore the option prices obtainable from this model should better compensate for the risk involved in entering an option contract. Furthermore, possible reformulations of Borland's model in a realm of semmimartingales were suggested.

References

[1] D. Applebaum. Lévy processes and stochastic calculus. *Cambridge University Press*, 2009.

[2] T. Bjork. Arbitrage theory in continuous time, 3rd edition. *Oxford University Press*, 2009.

[3] L. Borland. A theory of non-gaussian option pricing. *Quantitative Finance*, 2:415–431, 2002.

[4] R. Cont and P. Tankov. Financial modelling with jump processes. *A CRC Press Company*, 2004.

[5] B. Grigelionis. Student's t-distribution and related stochastic processes. *Springer*, 2013.

[6] F. C. Klebaner. Introduction to stochastic calculus with applications, 2nd edition. *Imperial College Press*, 2005.

[7] A. Shiryaev and O. E. Barndorff-Nielsen. Change of time and change of measure, 2nd edition. *World Scientific Publishing*, 2015.

[8] A. Shiryaev and J. Jacod. Limit theorems for stochastic processes, 2nd edition. *Springer*, 2003.

[9] P. Wilmott. On quantitative finance 2nd edition. *John Wiley & Sons*, 2006.

Quantification of Preferences for Markov Decision Processes*

Marko Ruman

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

advisors:

Tatiana Valentine Guy, Department of Adaptive Systems
Institute of Information Theory and Automation, CAS

Miroslav Kárný, Department of Adaptive Systems
Institute of Information Theory and Automation, CAS

Abstract. Recently, a preference elicitation for fully probabilistic design (FPD) of decision strategies has been proposed. The adopted FPD framework converts the preference elicitation to constructing an ideal probability. This ideal probability descries the desired closed-loop behaviour formed by all involved random variables. The choice of this ideal closed-loop model has been reduced to a minimisation task over a set of prospective ideals. The possibility that this set may be empty was left aside. Usual internal inconsistencies of the formulated preferences make this case frequent. The current paper solves with this problem. It applies the solution to Markov decision processes with discrete-valued closed-loop behaviours, which are quite vulnerable to inconsistencies. The result is broadly applicable as FPD is provably a proper dense extension of standard Bayesian decision making.

Keywords: Fully Probabilistic Design, Preference Elicitation, Markov Decision Processes, Kullback Leibler Divergence

Abstrakt. Tato práce navazuje na nedávno navrhnutou metodu konstrukce preferencí agenta v rozhodovacích úlohách s plně pravdě podobnostním návrhem (PPN) rozhodovacích strategií. Použití PPN převádí úlohu o konstrukci preferencí na konstruování ideální pravdepodobnostní funkce chování systému. Výběr této ideální pravdepodobnostní funkce byl převeden na maximalizační úlohu na množině přípustných idéalních pravdepodobnostních funkcí. V praktických úlohách se však často stáva, že množina přípustných idéalních pravdepodobnostních funkcí je prázdná. Tento případ je v článku vyřešen a řešení je aplikováno na Markovské rozhodovací úlohy s diskrétními stavy. Výsledky mají široké možnosti aplikace jelikož PPN je hustým rozšířením Bayesovské teorie rozhodování.

Klíčová slova: plně pravdepodobnostní návrh, konstrukce preferencí, Markovské rozhodovací procesy, Kullback Leibler divergence

Full paper: M. Ruman, T. V. Guy, M. Kárný. *Preference Elicitation for Markov Decision Processes within Fully Probabilistic Design Framework.* submitted to IEEE Transactions on Cybernetics.

*The research was supported by MSMT LTC18075 and by EU-COST Action CA16228.

Periods of Multidimensional Continued Fractions*

Hanka Řada

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

advisor: Štěpán Starosta, Department of Applied Mathematics
Faculty of Information Technology, CTU in Prague

Abstract. A classical continued fraction (CF) of a number x is periodic if and only if the number x is quadratic irrational. In the case of multidimensional continued fractions (MCFs), the situation is much more complicated. Firstly, the generalisation of classical CFs to more dimensions is unambigous. Secondly, it is known that if a vectorial n-dimensional CF algorithm (the algorithm that can be written as a multiplication of some integer matrices) of a vector $\vec{v}$ is periodic, then the elements of $\vec{v}$ are algebraic numbers of degree at most n. But it seems probable that the other direction does not hold in any of the well-known algorithms. In fact, there are only a few general results dealing with the problem which vectors have periodic MCF algorithm.

We present a new approach to this problem. We construct transducers for matrix Möbius transformations (transformations that can be written as a multiplication by an integer matrix) of Brun MCFs similar to the transducers of G. N. Raney [5]. Using these transducers, we investigate the periodicity of the transformed MCF from the MCF of the vector before the transformation and the periodicity of Brun MCFs in general.

Keywords: Multidimensional continued fractions, periodicity, algebraic degree, transducers

Abstrakt. Řetězový zlomek čísla x je periodický právě tehdy, když x je kvadratické iracionální číslo. V případě vícerozměrných řetězových zlomků (MCF) je však situace podstatně složitější. Za prvé neexistuje jednoznačné zobecnění řetězových zlomků pro více dimenzí. Za druhé je dokázáno, že pokud je vektorový n-dimenzionální MCF algoritmus (takový, který lze zapsat jako násobení celočíselných matic) vektoru $\vec{v}$ periodický, pak složky vektoru $\vec{v}$ jsou algebraická čísla, jejichž stupeň je maximálně n. Avšak opačná implikace velmi pravděpodobně neplatí u žádného z obecně známých algoritmů. Celkově existuje pouze velmi málo obecných výsledků zabývajících se otázkou, které vektory mají periodický MCF.

My zde představujeme nový přístup k tomuto problému. Konstruujeme převodníky pro maticové Möbiovy transformace (transformace, které lze zapsat jako násobní celočíselnou maticí) v Brunově MCF algoritmu, které jsou podobné Raneyho převodníkům [5] pro klasické řetězové zlomky. Pomocí těchto převodníků zkoumáme periodicitu transformovaného MCF pomocí periodicity MCF před transformací a také periodicitu Brunových MCF obecně.

Klíčová slova: Vícerozměrné řetězové zlomky, periodicita, algebraický stupeň, převodníky

*The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic, project no. CZ.02.1.01/0.0/0.0/16_019/0000778. H.Ř. acknowledges support by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/193/OHK4/3T/14. The computer experiments were done using the computer algebra system SageMath [9].

1 Introduction

The continued fraction algorithm is an iterative process, using which we can represent real numbers as a sequence of integers. It is based on Euclidean algorithm and it has many interesting properties. One of these properties is, that the continued fraction representation of a number is periodic if and only if the number is quadratic irrational (it is a root of a quadratic equation).

In 1839 Hermite asked Jacobi, if there is an algorithm which would detect the algebraic degree of a number (the minimal degree of a polynomial which has as a root the given number) also for other algebraic numbers (the numbers that can be written as a root of a polynomial). This gave birth to the multidimensional continued fractions (MCFs) which are a generalisation of the continued fractions to higher dimensions.

As the last nearly two hundert years showed, the Hermite's question is quite hard. The first big problem is, that there is no straightforward generalisation of the Euclidean algorithm to higher dimension. Therefore, there were created also many MCF algorithms. The second problem is, that for any of the well-known algorithms, there is no convincing proof of which numbers have periodic representation in the algorithm. Although there were many attempts to solve this problem (for example the numerous attempts of Leon Bernstein and many others), there are still only partially results.

We will study the Brun MCF algorithm and we will show some transducers using which we can compute Brun representation of a matrix Möbius transformation of a given vector (a multiplication of the vector by an integer matrix) directly from the Brun representation of this vector before the transformation. Using this transducers we introduce a concept that helps to detect the period of the result of such a transformation.

Moreover, we hope, that using this concept, we will be able to detect, if the result of a given matrix Möbius transformation is periodic or not. Or at least, that we find some vector whose elements create a base of a cubic number field and which does not have a periodic Brun expansion.

2 Classical continued fractions

Multidimensional continued fractions (MCFs) are generalisations of the classical continued fractions (CFs). Since all MCFs are based on the principle of classical continued fractions, to ease the understanding, we first introduce the classical CFs and some of their properties.

A step of *a continued fraction (CF) algorithm* (in a matrix form) is given by

$$\begin{cases} (x^i, y^i)^T \to (x^i - y^i, y^i)^T = (x^{i+1}, y^{i+1})^T, & \text{if } x^i \geq y^i \\ (x^i, y^i)^T \to (x^i, y^i - x^i)^T = (x^{i+1}, y^{i+1})^T, & \text{if } x^i < y^i. \end{cases}$$

Moreover let:

$$M_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \; M_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

Then we have:

$$\begin{pmatrix} x^i \\ y^i \end{pmatrix} = \begin{cases} M_1 \begin{pmatrix} x^{i+1} \\ y^{i+1} \end{pmatrix} & \text{if } x^i \geq y^i \\ \\ M_2 \begin{pmatrix} x^{i+1} \\ y^{i+1} \end{pmatrix} & \text{if } x^i < y^i. \end{cases}$$

Let $m \in \mathbb{N}$ and $\begin{pmatrix} x^\ell \\ y^\ell \end{pmatrix} \in \mathbb{R}_+^2$ for all $0 \leq \ell < m$. Then the first m steps of the algorithm can be written in the following way:

$$\begin{pmatrix} x^0 \\ y^0 \end{pmatrix} = M_{i_0} M_{i_1} \dots M_{i_{m-1}} \begin{pmatrix} x^m \\ y^m \end{pmatrix} = M_1^{a_0} M_2^{a_1} \dots M_{j_k}^{a_k} M_{j_{k+1}}^{\widehat{a_{k+1}}} \begin{pmatrix} x^m \\ y^m \end{pmatrix},$$

where $0 \leq k < m$ and $i_\ell, j_k, j_{k+1} \in \{1, 2\}$ for all $0 \leq \ell < m$ and $j_k \neq j_{k+1}$, $a_0 \in \mathbb{N}, a_\ell, \widehat{a_{k+1}} \in \mathbb{Z}_+$. Therefore the continued fraction representation of $\frac{x^0}{y^0}$ is $[a_0, a_1, a_2, \dots]$. This corresponds to the fact that:

$$\frac{x^0}{y^0} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots}}}.$$

Continued fractions have many interesting properties. In what follows, we sum up the properties, that we will investigate also in the multidimensional case.

2.1 Selected properties of classical CFs

Continued fractions can be used to determine some algebraic properties of a given number. More precisely, using the continued fraction algorithm, we can determinate, whether the number has algebraic degree equal to one, two or greater than two.

For $\frac{x^0}{y^0} \in \mathbb{Q}$, the number of steps in the CF algorithm is finite. For $\frac{x^0}{y^0} \in \mathbb{R} \setminus \mathbb{Q}$, the algorithm does not terminate, for arbitrary m the product of matrices M_1, M_2 corresponding to the first m steps in the CF algorithm is unique and determines the continued fraction of the number $\frac{x^0}{y^0}$.

Moreover, the CF algorithm of a number x is periodic if and only if x is a quadratic number (is the root of a quadratic polynomial and is irrational).

Example 1.

$$\begin{pmatrix} \sqrt{2} \\ 1 \end{pmatrix} = M_1 (M_2^2 M_1^2)^m \begin{pmatrix} \sqrt{2} \\ 1 \end{pmatrix}, \quad \forall m \in \mathbb{N}$$

Concerning these properties of classical CFs, there arises a natural question. Is there an analogue of the CF algorithm which would provide a periodic representation for every vector of the form $(x_1, x_2, \dots, x_n)^T \in \mathbb{R}_+^n$, where $x_1, x_2, \dots, x_n$ are numbers from the same number field of degree n?

Trying to answer this question, there were created many multidimensional continued fraction algorithms (MCFs). We will focus only on the vectorial ones (the algorithms that can be written as a matrix multiplications).

3 Multidimensional continued fractions

The Euclidean algorithm, used in the classical CFs, is not extensible in a straightforward manner to work with two and more numbers simultaneously. Because of that, there are many commonly used MCF algorithms. The most well-know ones are the Jacobi-Perron algorithm (introduced by Jacobi and generalized by Perron in [3]), Poincaré algorithm [4], Brun algorithm [2], Selmer algorithm [8] and Fully subtractive algorithm [6]. For a good overview of the known MCF algorithms see [1] and [7]. We decided to start with the investigation of the Brun algorithm in dimension $n = 3$ as it is one of the most simple ones (for classical CF we have $n = 2$).

3.1 Brun algorithm

Definition 2 (Brun algorithm [2] (1920)). *A step of the* Brun algorithm *is given by*

$$(x_1, x_2, x_3)^T \mapsto (x_1, x_2, x_3 - x_2)^T$$

for $0 \leq x_1 \leq x_2 \leq x_3$ *or analogously for other permutations of the vector* $(x_1, x_2, x_3)^T$.

We can again write the algorithm in the matrix form. In this form one step of the Brun algorithm for $0 \leq x_1 \leq x_2 \leq x_3$ is

$$(x_1, x_2, x_3)^T = B_4(x_1, x_2, x_3 - x_2)^T,$$

where $B_4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$. The matrices corresponding to the other permutations of (x_1, x_2, x_3) are $B_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, B_3 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B_5 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B_6 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$.

Brun in [2] also proved that for the dimension $n = 3$ every expansion in the Brun algorithm corresponds to a uniquelly determined vector (up to a scalar multiplication and permtuation of elements).

In what follows, we want to investigate the periodicity of the Brun algorithm. Therefore we start with the summary of what is known in this field.

3.2 Properties of MCFs and of the Brun algorithm

The following theorem holds for all (vectorial) MCF algorithms.

Theorem 3. *Suppose the MCF expansion of a vector* $(x_1, x_2, \ldots, x_n)^T$ *is periodic and let*

$$(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})^T = M(x_1^{(k+m)}, x_2^{(k+m)}, \ldots, x_n^{(k+m)})^T =$$
$$= \lambda M(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})^T$$

for some $k, m \in \mathbb{N}, m \neq 0$ *and* $\lambda \in \mathbb{R}$ *(*M *is the matrix of its periodic part and* λ *the eigenvalue of* M*). Then:*

- λ *is an algebraic unit of degree at most* n.
- *If the degree of* λ *equals* n, *then the numbers* $1, \frac{x_2}{x_1}, \frac{x_3}{x_1}, \ldots, \frac{x_n}{x_1}$ *constitute a basis of the number field* $\mathbb{Q}(\lambda)$.

The problem is that if $\deg(\lambda) \leq n-1$, we can have $\frac{x_j}{x_1} \notin \mathbb{Q}(\lambda)$.
For the Brun algorithm and $n = 3$ we know a little bit more.

Theorem 4 (Brun [2]). *The Brun algorithm for* $n = 3$ *detects rational dependence.*

This in other words means that the Brun algorithm for rationally dependent numbers terminates. Unfortunatelly, an analogue of this theorem does not hold even for $n = 4$.

For all of the MCF algorithms there still remain many open questions. A very important such a question is, if any of the well-known MCF algorithms can detect the algebraic degree of a number. In other words if in any of this algorithms for dimension $n \geq 3$ holds: a vector $x = (x_1, x_2, \ldots, x_n)$ has a periodic expansion if and only if $1, \frac{x_2}{x_1}, \frac{x_3}{x_1}, \ldots, \frac{x_n}{x_1}$ constitute a basis of a number field of degree n. This is not known even for $n = 3$.

The problem is, that it is really hard to say that something does not have periodic expansion (we can easily find some vector that does not have a periodic expansion in first $100, 1000, \ldots$ steps but we do not see what happens in the infinity).

Our goal is to solve, at least partially, the question, which numbers have periodic expansion in some of the well-known algorithms. For this purpose we try to investigate the MCF representation of a vector $y = Mx$, where M is a nonnegative integer matrix of the right size, compared with the MCF representation of a vector x. We again describe our methods firstly in the case of classical CFs.

4 Matrix Möbius transformations and transducers

The transformations we will look on are matrix Möbius transformations.

4.1 Matrix Möbius transformation and Raney's transducers for classical CFs

Definition 5. *Let* $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{N}^{2,2}$ *and* $\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2$.
The function $h : \mathbb{R}^2 \to \mathbb{R}^2$ *determined by*

$$h_M \begin{pmatrix} x \\ y \end{pmatrix} = M \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}$$

is the matrix Möbius transformation (MMT) *associated with the matrix* M.

For us is one of the key properties of this transformation, that the number $\frac{x}{y}$ is quadratic if and only if the number $\frac{\left(h_M \begin{pmatrix} x \\ y \end{pmatrix}\right)_1}{\left(h_M \begin{pmatrix} x \\ y \end{pmatrix}\right)_2}$ is quadratic.

For determining the CF representation of a number after a Möbius transformation we will use some sort of transducers. Before we introduce them, we will need another definition.

Definition 6. *Given $n > 0$, let $\mathcal{DB}_n$ denote the set of matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{N}^{2,2}$ satisfying:*

$$ad - bc = n$$
$$a > b, c \quad \text{and} \quad d > b, c.$$

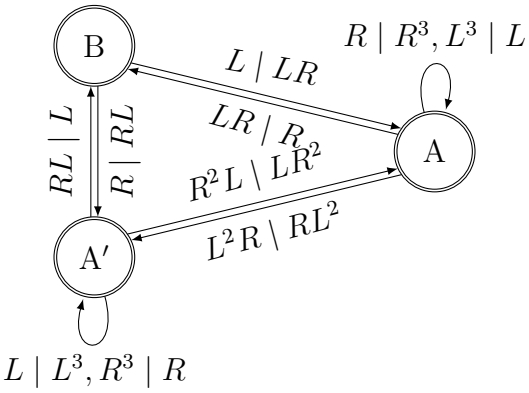
The shortcut $\mathcal{DB}$ here stand for 'doubly-balanced'. A little bit informally we can also say that $\mathcal{DB}$ is a set of matrices of the form:

$$\begin{pmatrix} a & > & b \\ \vee & & \wedge \\ c & < & d \end{pmatrix}.$$

For example $\begin{pmatrix} 4 & 1 \\ 2 & 4 \end{pmatrix} \in DB_{14}$.

G. N. Raney in [5] showed that for positive quadratic numbers x, y and a matrix $M \in \mathcal{DB}_n$ there exists a finete state transducer depending only on n, denoted $\mathcal{RT}_n$, that can be used to determine the matrix CF representation of $h_M \begin{pmatrix} x \\ y \end{pmatrix}$. Namely, the set of states of the transducer T_n is the set DB_n, the input word of this transducer is the matrix CF representation of $\begin{pmatrix} x \\ y \end{pmatrix}$, the initial state is M and the output word is the matrix representation of $h_M \begin{pmatrix} x \\ y \end{pmatrix}$. Moreover, Raney proved that the graph of the transducer $\mathcal{RT}_n$ is strongly connected.

In Figure 1 we can see the transducer $\mathcal{RT}_3$ and in Table 1 we can see the table of the transitions in this transducer.



$$\mathrm{A} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$
$$\mathrm{B} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$
$$\mathrm{A}' = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

Figure 1: Transducer $\mathcal{RT}_3$.

We try to use a generalisation of the matrix Möbius transformation and the transducers for that transformation for the MCFs. Therefore, we start by introducing, how this generalisation works.

	$\mathrm{A}=\begin{pmatrix}3&0\\0&1\end{pmatrix}$	$\mathrm{B}=\begin{pmatrix}2&1\\1&2\end{pmatrix}$	$\mathrm{A}'=\begin{pmatrix}1&0\\0&3\end{pmatrix}$
$\begin{pmatrix}3&0\\0&1\end{pmatrix}$	$R\|R^3, L^3\|L$	$LR\|R$	$L^2R\|RL^2$
$\begin{pmatrix}2&1\\1&2\end{pmatrix}$	$L\|LR$		$R\|RL$
$\begin{pmatrix}1&0\\0&3\end{pmatrix}$	$R^2L\|LR^2$	$RL\|L$	$L\|L^3, R^3\|R$

Table 1: Table of transitions in transducer $\mathcal{RT}_3$

4.2 Matrix Möbius transformations and transducers for MCFs

Definition 7. *Let* $M=\begin{pmatrix}a&b&c\\d&e&f\\g&h&i\end{pmatrix}\in\mathbb{N}^{3,3}$ *and* $\begin{pmatrix}x\\y\\z\end{pmatrix}\in\mathbb{R}^3$.

The function $h:\mathbb{R}^3\to\mathbb{R}^3$ *determined by*

$$h_M\begin{pmatrix}x\\y\\z\end{pmatrix}=M\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}ax+by+cz\\dx+ey+fz\\gx+hy+iz\end{pmatrix}.$$

is the matrix Möbius transformation (MMT) *associated with the matrix* M.

Analogously to the Raney's transducers we can construct transducers for performing matrix Möbius transformations also for the MCFs. The problem is, that in the case of MCFs the transducers do not have some of the key properties of the Raney's transducers. Namely, the transducers have infinitely many states. Moreover, it can happen that both the input and output words are eventually periodic and the transitions taken in the transducer are **not eventually in a cycle**. For that reason we introduce the concept of pattern transducers.

4.3 Pattern transducers for the Brun algorithm

For $Q\in\mathbb{Q}^{6,3}$ we set

$$\mathcal{P}_Q=\left\{A\in\mathbb{N}^{3,3}\colon A=\begin{pmatrix}x&y&z\\(w)_1&(w)_2&(w)_3\\(w)_4&(w)_5&(w)_6\end{pmatrix}\ \begin{matrix}\text{with } w=Q\begin{pmatrix}x\\y\\z\end{pmatrix}\\ \text{for some } x,y,z\in\mathbb{N}\end{matrix}\right\}.$$

Conjecture 8. *Let* $M,N\in\mathbb{N}^{3,3}, \det M, \det N>0$ *and* $NM\in\mathbb{Z}_+^{3,3}$. *There exists* $Q\in\mathbb{Q}^{6,3}$ *such that* $\forall m\in\mathbb{N}$

$$NM^m\in\mathcal{P}_Q.$$

Notation: given M,N, we set $P_{N,M}=\mathcal{P}_Q$.

Definition 9. *If* $P,Q\in\mathcal{P}_Q$ *we say that* P *and* Q have the same pattern.

Let us comment on, why we will use this concept. Let $x \in \mathbb{R}$ and $x = N\overline{M}$ in the Brun algorithm (N, M are words over the alphabet $\{B_1, B_2, \ldots, B_6\}$). So, if the above conjecture holds, we know that if we take the matrix corresponding to the preperiod N and first $10, 100, 1000$ or every other natural number periods M of the Brun representation of x we get a matrix with the same pattern ($NM^m \in P_{N,M}$ for all m). Moreover, in almost all cases, we need to compute only the matrices N, NM, NM^2 for determining the functions $(w)_1, (w)_2, \ldots, (w)_6$ from the definition of $\mathcal{P}_Q = P_{N,M}$.

For the case $N = \mathrm{Id}$ we can say even more.

Proposition 10.

$$P_{\mathrm{Id},M} = \mathcal{P}_Q$$

for some $Q = \begin{pmatrix} 0 & b_1 & c_1 \\ 1 & b_2 & c_2 \\ 0 & b_3 & c_3 \\ 0 & c_1 & c_4 \\ 0 & c_2 & c_5 \\ 1 & c_3 & c_6 \end{pmatrix} \in \mathbb{Q}^{6,3}$.

In other words, given $M \in \mathbb{Z}_+^{3,3}$ there exist $b_1, b_2, b_3, c_1, c_2, c_3, c_4, c_5, c_6 \in \mathbb{Q}$ such that $\forall m \in \mathbb{N}$

$$M^m = \begin{pmatrix} x & y & z \\ b_1 y + c_1 z & x + b_2 y + c_2 z & b_3 y + c_3 z \\ c_1 y + c_4 z & c_2 y + c_5 z & x + c_3 y + c_6 z \end{pmatrix}$$

for some $x, y, z \in \mathbb{Q}$.

We now use the concept of pattern matrices for constructing another transducers for matrix Möbius transformations.

Definition 11. *Let T be a MMT for an input word which has the matrix of preperiod equal to N and the matrix of period equal to M. Moreover, let $S_1 S_2 \ldots$, where $\forall i \in \mathbb{N}$, $S_i \in \{B_1, B_2, \ldots, B_6\}$, be the output word of this transformation and $S_0 = \mathrm{Id}$.*

The pattern transducer of T, N, M *is the transducer with states*

$$P_{TN,M}, P_{S_1^{-1}TN,M}, P_{S_2^{-1}S_1^{-1}TN,M}, \ldots$$

and transitions $P_{S_{i-1}^{-1}\ldots S_1^{-1}TN,M} \xrightarrow{M|S_i} P_{S_i^{-1}\ldots S_1^{-1}TN,M}$ *with* $i \in \mathbb{Z}_+$.

The calculation follows this walk:

$$P_{TN,M} \overset{M|S_1}{\to} P_{S_1^{-1}TN,M} \overset{M|S_2}{\to} P_{S_2^{-1}S_1^{-1}TN,M} \cdots$$

Let us comment a little bit more on, how the pattern transducers work. The pattern tranducers are created from the generalisation of Raney's tranducers. However, in the case of pattern tranducers, we in fact study, what happens if we read the preperiod and first $10, 100, 1000$ or infinitely many (or every other natural number) of periods of the input word before emitting the first part of the output word. And for that reason, we use the classes of the same pattern.

The following conjecture shows why is this concept useful.

Conjecture 12. *Let S be the matrix of the period of the output word of the MMT associated with the matrix T and the input word $NMMMMM\ldots$. Let $P_{N',M'}$ be the state in which the emission of the output word is already in its periodic part.*

Then $\forall R \in P_{N',M'}$ we have

$$S^{-1}R \in P_{N',M'}$$

and therefore also $\forall Q \in \mathcal{P}_{\mathrm{Id},S}$

$$QR \in P_{N',M'}.$$

We now explain, why is this conjecture so useful. If it holds then we know that, if we go through the pattern tranducer, we get in a cycle after emitting the period of the output word for the first time. Moreover, we can try to find if we are already emitting the period of the output word, in every state of the pattern tranducer. It works like this. We take the class $P_{S_{i-1}^{-1}\ldots S_1^{-1}TN,M}$ and find a class $\mathcal{P}_Q$ such that for all $P \in \mathcal{P}_Q$ and for all $R \in P_{S_{i-1}^{-1}\ldots S_1^{-1}TN,M}$ we have $PR \in P_{S_{i-1}^{-1}\ldots S_1^{-1}TN,M}$. Then we try to find, if a Brun matrix $V \in \mathcal{P}_Q$ exists (Brun matrix is a matrix that corresponds to some expansion in the Brun algorithm). If there is such V it is the matrix of the period (or some multiple of the period) of the output word.

Even more interesting is the idea to use this conjecture to find that we are not yet emitting the period of the output word. If we would be able to determine that, then it is possible that we would be able to find a vector which elements constitute a basis of some cubic number field and which has an aperiodic expansion in the Brun algorithm.

References

[1] A. J. Brentjes. *Multi-dimensional continued fraction algorithms.* MC Tracts (1981).

[2] V. Brun. *En generalisation av Kjedebroken.* na, (1920).

[3] O. Perron. *Grundlagen für eine theorie des jacobischen kettenalgorithmus.* Math. Ann. (1907).

[4] H. Poincaré. *Sur une généralisation des fractions continues.* CR Acad. Sci. Paris. Ser **1** (1884), 1014–1016.

[5] G. N. Raney. *On continued fractions and finite automata.* Mathematische Annalen **206** (1973), 265–283.

[6] F. Schweiger. *Invariant measures for fully subtractive algorithms.* Anz. Österreich. Akad. Wiss. Math.-Natur. Kl **131** (1995), 25–30.

[7] F. Schweiger. *Multidimensional continued fractions.* Oxford University Press on Demand, (2000).

[8] E. S. SELMER. *Om flerdimensjonal kjedebrøk.* Nordisk Matematisk Tidskrift (1961), 37–43.

[9] The Sage Developers. *SageMath, the Sage Mathematics Software System*, (2018). http://www.sagemath.org.

Log-Anharmonic Oscillator and Its Large-N Solution*

Iveta Semorádová

4th year of PGS, email: Iveta.Semoradova@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. The large-N expansion technique is tested via an anomalous, soft-core potential which admits the tunneling through its central barrier. The precision of the approximation is found sensitive to the asymptotic component of the interaction. Once chosen in the most common harmonic-oscillator form, and once complemented by the short range part represented by the general power-law anharmonicity $\sim |x|^\alpha$, we found that the latter power-law spike may be well approximated by an elementary logarithmic function, in the limit of the smallest $\alpha \to 0$ at least. In such a model, the large-N method is found applicable and offering still an efficient and user-friendly method of the solution of the Schrödinger equation.

Keywords: Confining interactions, soft central repulsion, large-N expansion method, logarithmic anharmonicity.

Abstrakt. Technika rozvoje v mocninnách $1/N$ je testována pomocí anomálního potenciálu, který umožňuje tunelování skrz centrální bariéru. Přesnost aproximace je citlivá vzhledem k asymptotickému členu interakce. Vybereme-li nejběžnější formu odpovídající haronickému oscilátoru a doplníme-li ji krátkodosahovou mocninnou anharmonickou částí s malým exponentem $\sim |x|^\alpha$, zjistíme, že i tento člen může být dobře aproximován pomocí logaritmické funkce, přinejmenším v limitě $\alpha \to 0$. V takovémto modelu je metoda rozvoje v mocninnách $1/N$ dobře aplikovatelná a nabízí efektivní a uživatelsky přívětivou metodu řešení Schrödingerovy rovnice.

Klíčová slova: Věznící interakce, slabé centrální odpuzování, metoda rozvoje v mocninnách $1/N$, logaritmická anharmonicita.

Full paper: M. Znojil, I. Semorádová. *Log-anharmonic oscillator and its large-N solution.* Modern Physics Letters A **33** (2018), 1850223.

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS19/183/OHK4/3T/14.

State Transfer Algorithm Based on Discrete-Time Quantum Walks*

Stanislav Skoupý

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

advisor: Martin Štefaňák, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. In this article we focus on state transfer algorithm based on discrete-time quantum walks. We introduce new version of this algorithm which we compare to the original version. We present 2 conditions to achieved perfect state transfer with new algorithm. These conditions are for search algorithm on which the state transfer algorithm is based on. We show that on example of M-partite graph the new algorithm performs better than the original one.

Keywords: quantum walk, search algorithm, state transfer

Abstrakt. V tomto článku se zaměříme na algoritmus pro přenos stavu založený na kvantových procházkách v diskrétním čase. Představíme novou verzi tohoto algoritmu, kterou porovnáme s jeho originální verzí. Představíme 2 podmínky vyhledávací algoritmu, na kterém se přenos stavu založen, při jejichž splnění nový algoritmus dosáhne úplného přenosu stavu. Ukážeme, že na M-partitního grafu nový algoritmus funguje lépe než originál.

Klíčová slova: kvantová procházka, vyhledávací algoritmus, přenos stavu

1 Introduction

Quantum walks are important part of quantum computer science. Search algorithm and state transfer algorithm that can be implemented using framework of quantum walk. Using model of discrete time quantum walks with coins we propose new state transfer algorithm based closely on the search algorithm than the original state transfer algorithm. We also present example of graph where new algorithm performs better than the original.

In the first section we describe the general scheme of the search algorithm, the original state transfer algorithm and the new state transfer algorithm. Then we discuss the differences of the two transfer algorithm. In the last section we present example of complete M-partite graph, where new algorithm outperforms the original algorithm.

2 General scheme

In this section we present first the general scheme of the search algorithm and original state transfer algorithm based on the discrete-time quantum walks and. These algorithms

*This work was supported from Student Grant Competition of Czech Technical University in Prague under Grant SGS19/186/OHK4/3T/14, from GAČR under Grant No. 17-00804S and project CAAS.

are based on work in [1], [2] and [3]. Then we introduce new scheme of the state transfer algorithm. Advantages and disadvantages of new state transfer algorithm are discussed in the following section.

Before we describe the algorithms, we have to first describe the model of discrete time quantum walks with coins. Having a graph $G = (V, E)$ the corresponding Hilbert space of the walk $\mathcal{H}_G$ is spanned by basis $\{|v, w\rangle | \forall v, w \in V : \{v, w\} \in E\}$ where the first index v describe the position of the walker and the second index describe w describes the direction of the walker. Movement of the walker is achieved by application of shift operator $\hat{S}$ which is defined in following way

$$\hat{S}|v, w\rangle = |w, v\rangle. \tag{1}$$

Using only shift operator the evolution of one step of the walk is trivial and two steps are equal to identity. Hence, the coin operator $\hat{C}$ is additionally applied at every step. The coin operator acts locally at coin subspace at each vertex v that is spanned by states $\{|v, w\rangle | \forall w \in V : \{v, w\} \in E\}$. Then the evolution operator $\hat{U}$ of one step of the walk consist of application of the coin operator followed by application of the shift operator, i.e. $\hat{U} = \hat{S}\hat{C}$

The main idea of search and state transfer algorithm is applying one local coin operator to marked vertices and different local coin operator to other vertices of the graph. The local coin operator that we use on non-marked vertices is known as Grover operator [4]

$$\hat{G}_v = -\hat{I} + 2|\psi_v\rangle\langle\psi_v| \tag{2}$$

where $|\psi_v\rangle$ is equal superposition of all direction of the coin space at vertex v. At the marked vertices often used coins are simple phase shift by π or the Grover operator followed by phase shift by π. The coin operator of the search algorithm with one marked vertex m, which we want to find, reads

$$\hat{C}_m = \sum_{\substack{v \in V \\ v \neq m}} \hat{G}_v \hat{P}_v + \hat{C}_m \hat{P}_m \tag{3}$$

where $\hat{C}_m$ is the coin operator at the marked vertex and $\hat{P}_v$ is projector on the coin subspace at the vertex v. Using coins operator (3) we get the evolution operator the search algorithm $\hat{U}_m$. We can finally introduce the steps of the search algorithms as follows:

1. Initialize the system in the superposition of all basis states

$$|init\rangle = \frac{1}{\sqrt{2|E|}} \sum_{v \in V} \sum_{\substack{w \\ \{v,w\} \in E}} |v, w\rangle. \tag{4}$$

2. Apply the evolution operator $\hat{U}_m$ T-times.

3. Measure the system.

The success probability and number of steps T depends on the structure and the size of the graph G.

In the case of state transfer algorithm we have 2 vertices instead one, sender and receiver. The original state transfer algorithms uses coin operator that applies the same local coin at both vertices at the same time. It has the following form

$$\hat{C}_{s,r} = \sum_{\substack{v \in V \\ v \neq s,r}} \hat{G}_v \hat{P}_v + \hat{C}_s \hat{P}_s + \hat{C}_r \hat{P}_r. \tag{5}$$

Having the state transfer coin operator the steps of original state transfer algorithm are following:

1. Initialize the system in the superposition of all directions at the sender vertex

$$|init\rangle = \frac{1}{\sqrt{d\,(s)}} \sum_{\substack{w \\ \{s,w\} \in E}} |s, w\rangle. \tag{6}$$

2. Apply the evolution operator $\hat{U}_{s,r}$ T'-times.

3. Measure the system.

The fidelity of the state transfer and the number of steps T' depends again on the size and structure of the graph.

We propose the slightly different approach to the search algorithm. Instead of using coin operator of state transfer (5) we use the coin operator in the search algorithm and swish the marked vertex from sender to receiver after half number of steps. Hence, the algorithm goes as follows

1. Initialize the system in the target state of the search algorithm if we searched for sender vertex $|s\rangle$.

2. Apply the evolution operator $\hat{U}_s$ T-times.

3. Apply the evolution operator $\hat{U}_r$ T-times.

4. Measure the system.

We see that the number of steps of this algorithm is $2T$, i.e. twice the number of steps of search algorithm with one marked vertex. As we show in the following section, lower boundary of the fidelity can be found using only results from the search algorithm with one marked vertex, which is not true for the original state transfer algorithm.

3 State transfer algorithms and lower boundary of fidelity

In this section we discuss the differences of two state transfer algorithms from previous section and we show how to derive the lower bound of fidelity of the state transfer if

the search algorithm fulfils two conditions. Let us start with disadvantage of new state transfer algorithm that is change of the evolution operator during the run of the algorithm whereas original state transfer algorithm uses one evolution operator during the whole run. This might complicate physical implementation of the algorithm. Also we have to calculate the target state of search algorithm for searched of the sender vertex and then initialise the system in this state while in the original algorithm the initial state is fixed to equal superposition of all direction at sender vertex. Advantage of new algorithm is that the fidelity of the transfer can be lower bounded using results from search algorithm. In the case, where the probability of the search algorithm goes to 1, fidelity of the state transfer goes to 1 as well, hence perfect state transfer is achieved. Graphs, where the search algorithm perform with probability close to one, are strong candidates where the perfect state transfer is achieved using original algorithm, but no general relation between success probability and fidelity has been found. In [5] and [6] we present examples where the perfect state transfer is achieved using original transfer algorithm and search algorithm succeed with probability close to 1 on those graphs. However, in last section we present example of graph where search algorithm works with probability close to 1 and original state transfer algorithm does not achieved perfect state transfer. At this case new algorithm outperforms the original one.

Now we introduce two condition for the search algorithm from which we derive the lower bound for fidelity of new transfer algorithm. First condition is related to probability of success of search algorithm. We suppose that there exist T such that this hold true

$$\left(\hat{U}_m\right)^T |init\rangle = \alpha_m |m\rangle + \epsilon_m |y_m\rangle \tag{7}$$

for $|\alpha_m|$ close to one and $|\epsilon_m| \ll 1$, where $|y_m\rangle$ is unit vector and $|m\rangle$ is target state, i.e. if the system is in this state success probability is exactly 1. Second condition describe periodicity of the walks and it reads

$$\left(\hat{U}_m\right)^{2T} |init\rangle = \beta_m |init\rangle + \delta_m |z_m\rangle \tag{8}$$

for $|\beta_m|$ close to one and $|\delta_m| \ll 1$, where $|z_m\rangle$ is unit vector. These two condition are quite natural for any search algorithms with success probability close to one.

Using both condition we write the final state of new state transfer algorithm in following manner

$$\begin{aligned}
\left(\hat{U}_r\right)^T \left(\hat{U}_s\right)^T |s\rangle &= \frac{1}{\alpha_s}\left(\hat{U}_r\right)^T \left(\hat{U}_s\right)^T \left(\left(\hat{U}_s\right)^T |init\rangle - \epsilon_s |y_s\rangle\right) = \\
&= \frac{1}{\alpha_s}\left(\hat{U}_r\right)^T \left(\left(\hat{U}_s\right)^{2T} |init\rangle - \epsilon_s \left(\hat{U}_s\right)^T |y_s\rangle\right) = \\
&= \frac{1}{\alpha_s}\left(\hat{U}_r\right)^T \left(\beta_s |init\rangle + \delta_s |z_s\rangle - \epsilon_s \left(\hat{U}_s\right)^T |y_s\rangle\right) = \\
&= \frac{\alpha_r}{\alpha_s}\beta_s |r\rangle + \frac{\beta_s}{\alpha_s}\epsilon_r |y_r\rangle + \frac{\delta_s}{\alpha_s}\left(\hat{U}_r\right)^T |z_s\rangle - \frac{\epsilon_s}{\alpha_s}\left(\hat{U}_r\right)^T \left(\hat{U}_s\right)^T |y_s\rangle
\end{aligned} \tag{9}$$

where we first use (7) for marked vertex $|s\rangle$, then we use (8) and finally we again use (7) for state $|r\rangle$. Now to approximate $\left|\left\langle r \left| \left(\hat{U}_r\right)^T \left(\hat{U}_s\right)^T \right| s \right\rangle\right|$, which is square root of

fidelity, we use following estimations

$$\begin{aligned} |\langle r|y_r\rangle| &\leq |||r\rangle||\,|||y_r\rangle|| = 1 \\ \left|\left\langle r\left|\left(\hat{U}_r\right)^T\right|z_s\right\rangle\right| &\leq |||r\rangle||\left|\left|\left(\hat{U}_r\right)^T|z_s\rangle\right|\right| = |||r\rangle||\,|||z_s\rangle|| = 1 \\ \left|\left\langle r\left|\left(\hat{U}_r\right)^T\left(\hat{U}_s\right)^T\right|y_s\right\rangle\right| &\leq |||r\rangle||\left|\left|\left(\hat{U}_r\right)^T\left(\hat{U}_s\right)^T|y_s\rangle\right|\right| = 1 \end{aligned} \tag{10}$$

which are simple results Cauchy–Schwarz inequality and unitary of evolution operator. Using (9) and (10) we get lower bound for square root of the fidelity which reads

$$\begin{aligned} \left|\left\langle r\left|\left(\hat{U}_r\right)^T\left(\hat{U}_s\right)^T\right|s\right\rangle\right| &\geq \frac{|\alpha_r|}{|\alpha_s|}|\beta_s| - \frac{|\beta_s|}{|\alpha_s|}|\epsilon_r|\,|\langle r|y_r\rangle| - \\ &- \frac{|\delta_s|}{|\alpha_s|}\left|\left\langle r\left|\left(\hat{U}_r\right)^T\right|z_s\right\rangle\right| - \frac{|\epsilon_s|}{|\alpha_s|}\left|\left\langle r\left|\left(\hat{U}_r\right)^T\left(\hat{U}_s\right)^T\right|y_s\right\rangle\right| \geq. \\ &\geq \frac{|\alpha_r|}{|\alpha_s|}|\beta_s| - \frac{|\beta_s|}{|\alpha_s|}|\epsilon_r| - \frac{|\delta_s|}{|\alpha_s|} - \frac{|\epsilon_s|}{|\alpha_s|} \end{aligned} \tag{11}$$

It is easy to see from (11) that if $|\alpha_s|$, $|\alpha_r|$ and $|\beta_s|$ are close to one and if $|\epsilon_s| \ll 1$, $|\epsilon_r| \ll 1$ and $|\delta_s| \ll 1$ then the fidelity of the transfer is close to one.

4 M-partite graph

In this section we show an example of graph where new state transfer algorithm outperforms the original one. This example is complete M-partite graph with one self loop at each vertex. M-partite graph is graph with sets of vertices V divided into M subsets where there are no edges between vertices in one subset. In complete M-partite graph each vertex is connected to all vertices in other subsets. We also limited ourself to the case where all subsets have the same size N, thus whole graph has MN vertices. In [7] is shown that search algorithm works on complete M-partite graph but not with probability close to 1. To improve the success probability we use the same trick that was used by Wong in [8] by adding self loop in each vertex. As we show later this achieved success probability close to one and also search algorithm fulfils conditions (7) and (8). Hence we get perfect state transfer using new algorithm. In contrast the fidelity of the original transfer algorithm does not reach even one half, see the Figure 1.

Now we prove that the search algorithm succeed with probability close to one. Let us

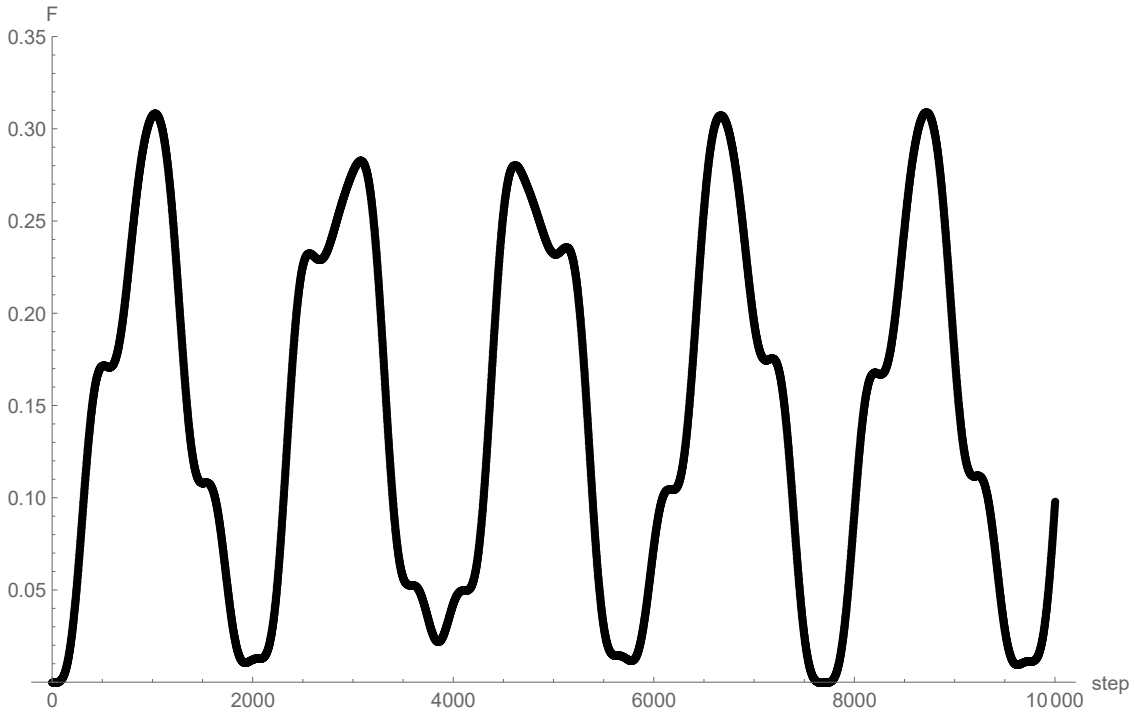


Figure 1: There is evolution of the fidelity $\mathcal{F}$ of the state transfer during 1000 steps complete M-partite graph with 100 parts each with 1000 vertices. Fidelity does not grow over 0.35.

introduce subspace which is spanned by following states

$$
\begin{aligned}
|\nu_1\rangle &= \frac{1}{\sqrt{(M-1)N}} \sum_{\alpha=2}^{M} \sum_{k=1}^{N} |1, m, \alpha, k\rangle \\
|\nu_2\rangle &= |1, m, 1, m\rangle \\
|\nu_3\rangle &= \frac{1}{\sqrt{(M-1)N(N-1)}} \sum_{j\neq m}^{N} \sum_{\alpha=2}^{M} \sum_{k=1}^{N} |1, j, \alpha, k\rangle \\
|\nu_4\rangle &= \frac{1}{\sqrt{(N-1)}} \sum_{j\neq m}^{N} |1, j, 1, j\rangle \\
|\nu_5\rangle &= \hat{S}|\nu_1\rangle = \frac{1}{\sqrt{(M-1)N}} \sum_{\alpha=2}^{M} \sum_{k=1}^{N} |\alpha, k, 1, m\rangle \\
|\nu_6\rangle &= \hat{S}|\nu_3\rangle = \frac{1}{\sqrt{(M-1)N(N-1)}} \sum_{j\neq m}^{N} \sum_{\alpha=2}^{M} \sum_{k=1}^{N} |\alpha, k, 1, j\rangle \\
|\nu_7\rangle &= \frac{1}{N\sqrt{(M-1)(M-2)}} \sum_{\substack{\alpha,\beta=2 \\ \beta\neq\alpha}}^{M} \sum_{j,k=1}^{N} |\alpha, j, \beta, k\rangle \\
|\nu_8\rangle &= \frac{1}{\sqrt{(M-1)N}} \sum_{\alpha=2}^{M} \sum_{k=1}^{N} |\alpha, k, \alpha, k\rangle
\end{aligned}
\qquad . \tag{12}
$$

where first 2 indexes in $|\alpha, k, \beta, j\rangle$ describe position and second pair of indexes describe direction, Greek letter label the part and Latin letters label position in that part. Marked vertex is in first part at vertex label by m. This subspace is invariant under evolution operator of the search algorithm and the initial state and target state lies in this subspace, thus we reduce the calculation to this subspace without loss of any information from the search algorithm. We introduce effective evolution operator which is evolution operator of the search in the basis of invariant subspace and it reads

$$U_{eff} = \begin{pmatrix} 0 & 0 & 0 & 0 & \frac{-N(M-1)+1}{N(M-1)+1} & \frac{2\sqrt{N-1}}{N(M-1)+1} & \frac{2\sqrt{(M-2)N}}{N(M-1)+1} & \frac{2}{N(M-1)+1} \\ -\frac{2\sqrt{(M-1)N}}{N(M-1)+1} & \frac{N(M-1)-1}{N(M-1)+1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{2\sqrt{N-1}}{N(M-1)+1} & -\frac{N(M-3)+3}{N(M-1)+1} & \frac{2\sqrt{(M-2)N(N-1)}}{N(M-1)+1} & \frac{2\sqrt{N-1}}{N(M-1)+1} \\ 0 & 0 & \frac{2\sqrt{(M-1)N}}{N(M-1)+1} & \frac{-N(M-1)+1}{N(M-1)+1} & 0 & 0 & 0 & 0 \\ \frac{-N(M-1)+1}{N(M-1)+1} & -\frac{2\sqrt{(M-1)N}}{N(M-1)+1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{N(M-1)-1}{N(M-1)+1} & \frac{2\sqrt{(M-1)N}}{N(M-1)+1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{2\sqrt{(M-2)N}}{N(M-1)+1} & \frac{2\sqrt{(M-2)N(N-1)}}{N(M-1)+1} & \frac{N(M-4)-1}{N(M-1)+1} & \frac{2\sqrt{(M-2)N}}{N(M-1)+1} \\ 0 & 0 & 0 & 0 & \frac{2}{N(M-1)+1} & \frac{2\sqrt{N-1}}{N(M-1)+1} & \frac{2\sqrt{(M-2)N}}{N(M-1)+1} & \frac{-N(M-1)+1}{N(M-1)+1} \end{pmatrix} \tag{13}$$

Spectrum of U_{eff} is composed of two pairs of conjugate eigenvalues, non degenerated eigenvalue 1 and three times degenerated eigenvalue -1. Eigenvectors are to complicated and long to contain in this article.

Numerical simulation show that the target state of the search algorithm is the state $|\nu_2\rangle$, i.e. the loop at the marked vertex. Using the calculation of the spectrum of (13) we get

$$|\langle \nu_2 | U^T | init \rangle|^2 = \frac{\left(\cos\left(\omega_1 T\right) - 1\right)^2}{4} - O\left(\frac{1}{M}\right) - O\left(\frac{1}{N}\right) \tag{14}$$

where ω is eigenphase of one pair of conjugate eigenvalues which reads

$$\omega_1 \quad = \arccos\left(\frac{N(M-2)+1+\sqrt{M^2N^2-6MN+4N+5}}{2N(M-1)+2}\right). \tag{15}$$

This result fulfils the first condition (7) for T closet integer to the number $\frac{\pi}{\omega_1}$.Also from (14) we get that the initial state of new transfer algorithm is state corresponding to the loop at sender vertex. The second condition is fulfil from following expression

$$|\langle init | U^T | init \rangle|^2 = \frac{\left(\cos\left(\omega_1 T\right) + 1\right)^2}{4} - O\left(\frac{1}{M}\right) - O\left(\frac{1}{N}\right). \tag{16}$$

We show that the search algorithm fulfils two conditions and hence new search algorithm achieves perfect state transfer on complete M-partite graph with one self loop at each vertex.

5 Conclusion

We introduce the new state transfer algorithm and we compared it to the original state transfer algorithm. We show its advantages and disadvantages. We prove that if the search algorithms fulfils two conditions (7) and (8) then the new state transfer algorithm achieves perfect state transfer. We use these conditions to show that on complete M-partite graph new algorithm perform better then the original one.

References

[1] N. Shenvi, J. Kempe, K. B. Whaley. *A quantum random walk search algorithm*, Phys. Rev. A 67, 052307 (2003)

[2] A. Ambainis, J. Kempe, A. Rivoch. *Coins make quantum walks faster*, Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (2005), p. 1099

[3] B. Hein, G. Tanner. *Wave communication across regular lattice*, Phys. Rev. Lett. 103, 260501 (2009)

[4] Lov K. Grover. *Quantum Mechanic helps in searching for a needle in a haystack*, Phys. Rev Lett. 78, 325 (1997)

[5] M. Štefaňák, S. Skoupý. *Perfect state transfer by means of discrete-time quantum walk search algorithms on highly symmetric graphs*, Phys. Rev. A 94, 022301 (2016)

[6] M. Štefaňák, S. Skoupý. *Perfect state transfer by means of discrete-time quantum walk on complete bipartite graphs*, Quantum Inf. Process. 16, 72 (2017)

[7] D. Reitzner, M. Hillery, E. Feldman, V. Bužek *Quantum searches on highly symmetric graphs*, Phys. Rev. A 79, 012323 (2009)

[8] T. G. Wong. *Grover Search with Lackadaisical Quantum Walks*, J.Phys. A 48, 435304 (2015)

Numerical Modelling of the Adsorption and Desorption of the Water Vapor in the Zeolite 13X Using a Two-Temperature Model and Mixed-Hybrid Finite Element Method Numerical Solver*

Tomáš Smejkal

3rd year of PGS, email: smejkto5@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jiří Mikyška, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The transition from fossil fuels to cleaner and renewable energy sources is currently one of the highest world priority. Solar energy is one of the most suitable choices for replacing fossil fuels. However, to use solar energy to its maximum potential, proper heat energy storage has to be designed. In recent years, various approaches for the heat energy storage has been proposed and studied. These approaches can be divided into three categories based on how the heat is stored: latent, sensible, or thermo-chemical. In this work, we are interested in the thermo-chemical heat energy storage using the zeolite, which is a crystalline aluminosilicate with a specific structure and a large internal surface area.

In this contribution, we present a new two-temperature mathematical model for a thermo-chemical energy storage based on the adsorption and desorption of the water vapor in zeolite 13X. The adsorption process is modelled using the Linear Driving Force (LDF) model and the Langmuir-Freundlich isotherms. A numerical solver based on the mixed-hybrid finite element method and operator splitting technique is proposed. Furthermore, we present a computational study of the charging and discharging processes of the thermo-chemical energy storage emphasising the behaviour of the two temperatures: fluid temperature and zeolite temperature.

Keywords: zeolite 13X; adsorption; desorption; LDF model; thermo-chemical energy storage; MHFEM; Langmuir-Freundlich isotherms; operator splitting

Abstrakt. Přechod z fosilních paliv na čistší a obnovitelné zdroje energie je v současné době jednou z nejvyšších světových priorit. Sluneční energie se jeví jako jedna z nejvhodnějších možností. Aby však bylo možné využít sluneční energii na maximální potenciál, musí být k dispozici vyhovující zařízení pro akumulaci této energie. V posledních letech byly navrženy a studovány různé přístupy k akumulaci tepelné energie. Tyto přístupy lze rozdělit do tří kategorií podle toho, jak je teplo ukládáno: latentní, citlivé nebo termo-chemické. V této práci se zajímáme o

*The work was supported by the project Thermal energy storage materials: thermophysical characteristics for the design of thermal batteries, project no. 17-08218S of the Czech Science Foundation, 2017-2019 and the project Application of advanced supercomputing methods in mathematical modeling of natural processes, grant no. SGS17/194/OHK4/3T/14 of the Grant Agency of the Czech Technical University in Prague, 2017-2019.

termo-chemickou akumulaci tepelné energie pomocí zeolitu, což je krystalický hlinito-křemičitan se specifickou strukturou a velkou vnitřní povrchovou plochou.

V tomto příspěvku prezentujeme nový dvou-teplotní matematický model pro ukládání tepelné energie založený na absorpci a desorpci vodních par na zrnech zeolitu 13X. Proces absopce je modelován pomocí LDF (Linear Driving Force) modelu a Langmuirových-Freundlichových isoterm. Numerický řešič je založen na smíšené hybridní metodě konečných objemů a technice rozdělení operátoru (operator splitting technique). Výpočetní studie pro nabíjecí a vybíjecí proces je prezentována. Provedeme diskuzi nad chováním dvou-teplotního modelu.

Klíčová slova: zeolit 13X; adsorpce; desorpce; LDF model; termo-chemické ukládání energie; MHFEM; Langmuirovy-Freundlichovy isotermy; rozdělení operátoru

Full paper: Smejkal T., Mikyška J., Fučík R., Numerical modelling of the adsorption and desorption of the water vapor in the zeolite 13X using a two-temperature model and mixed-hybrid finite element method numerical solver, submitted to International Journal of Heat & Mass Transfer (2019).

Dimensional Effects of Inter-Phase Mass Transfer on Attenuation of Structurally Trapped Gaseous Carbon Dioxide in Shallow Aquifers*

Jakub Solovský

4th year of PGS, email: jakub.solovsky@fjfi.cvut.cz
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. Based on experimental evidence and using mathematical modeling, inter-phase mass transfer processes of CO_2 exsolving from and dissolving into water in heterogeneous porous media are investigated under two fundamentally different flow conditions: in a quasi one dimensional vertical column and in a two-dimensional tank with a lateral background water flow, both at laboratory scale. In both cases, the CO_2 dissolved in water under a given overpressure is injected for a certain period at the bottom of the tank, exsolves, and migrates upwards. A layer of fine sand is present in the tanks designed to mimic geological scenarios of accumulation and trapping of exsolved CO_2 in shallow aquifers. Then, clean water is injected and the accumulated CO_2 is dissolved back into the flowing water. The study aims to point out the differences in the mass transfer processes between the quasi-1D and 2D cases using a mathematical model of two-phase compositional flow in heterogeneous porous media calibrated to the experimental datasets, and expose strategies that should be explored in the future research. Additionally, temperature variations observed during the 2D experiments allow for analysis of isothermal versus non-isothermal effects on the processes of multiphase CO_2 evolution. The mathematical model is discretized and solved using the mixed hybrid finite element method in 2D that allows for the simulation both advection- and diffusion-dominated processes accurately.

Keywords: compositional flow, two–phase flow, non–equilibrium mass transfer, kinetic mass transfer, gas exsolution, gas dissolution

Abstrakt. Na základě experimentálních dat a matematického modelu jsou procesy vývinu a rozpouštění CO_2 ve vodě v heterogenním porézním prostředí zkoumány ve dvou různých případech: ve kvazi-jednorozměrném vertikálním sloupci a ve dvourozměrném případě s prouděním

*The work reported in this paper was supported by the Czech Science Foundation project no. 17-06759, by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/194/OHK4/3T/14, and by the Centre of Advanced Applied Sciences project with the number: CZ.02.1.01/0.0/0.0/16-019/0000778. The authors would like to thank the U.S. Department of Energy's Office of Fossil Energy for providing funding for this research through the National Energy Technology Laboratory's CO_2 sequestration R&D Program. The support of Dr. Rajesh Pawar of LANL is gratefully acknowledged. National Science Foundation also deserves credit for providing funding for this research through award 1045282. The United States Geological Survey did not collect the data used, and did not perform the simulations presented, in this work. Any use of trade, product or firm names in this publication is for descriptive purposes only and does not imply endorsement by the United States Government.

vody celou oblastí. Oba případy jsou uvažovány v laboratorním měřítku. V obou případech je voda při daném přetlaku nasycena CO_2 a po danou dobu vtláčena do spodní části uvažované oblasti. V oblasti se CO_2 vyvíjí jako samostatná plynná fáze a stoupá vzhůru. V oblasti je přítomna vrstva jemného písku, která napodobuje situace v mělkých rezervoárech. V okolí této vrstvy dochází k akumulaci a zachytávání CO_2. Následně je do oblasti vtláčena čistá voda, do které se postupně rozpouští CO_2 nahromaděný v plynné fázi. Studie je zaměřená na zkoumání rozdílů v přestupu hmoty mezi kvazi-1D a 2D případem za použití matematického modelu dvoufázového kompozičního proudění v heterogenním porézním prostředí kalibrovaném na experimentální data. V této studii jsou také navrženy strategie pro další výzkum těchto procesů. V průběhu 2D experimentu navíc došlo ke změnám teploty. Tyto fluktuace umožnily zkoumat vliv změn teploty na rozpouštění a vývin CO_2. Matematická formulace problému je diskretizována a řešena za použití smíšené hybridní metody konečných prvků ve 2D. Tato metoda umožňuje přesné řešení jak advekčně- tak difuzně- dominantních úloh.

Klíčová slova: kompoziční proudění, dvoufázové proudění, nerovnovážný přestup hmoty, kinetický přestup hmoty, vývin plynu, rozpouštění plynu

Full paper: J. Solovský, R. Fučík, M. R. Plampin, T. H. Illangasekare, J. Mikyška. *Dimensional Effects of Inter-phase Mass Transfer on Attenuation of Structurally Trapped Gaseous Carbon Dioxide in Shallow Aquifers.* Submitted to Journal of Computational Physics (2019).

Emotional Anomaly Detection

Zuzana Szabová

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

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

Abstract. Knowledge of either single employees emotional state or collective mood is a key managerial task. In this paper we describe methods able to evaluate both individual and team mood state. Firstly we introduce metric that enables evaluating similarity between individuals or teams based on their mood state. Subsequent application of dimensionality reduction methods enabled visualisation of similarities in 2-D space. Finally, clustering algorithm were used to divide users into well distinguishable clusters and create artificial historical data for new individuals.

Keywords: t-SNE, Dimensionality reduction, Distribution similarity measures, Moods, anomaly detection

Abstrakt. Znalost jednak současného i dlouhodobého stavu, jak zaměstnanců, tak celého týmu je jednou z klíčových rolí každého manažera. V tomto příspěvku prezentujeme metody vedoucí k automatizovanému vyhodnocování stavu jednotlivců i kolektivu. Prvně je představena metrika, jež slouží k evaluaci podobnosti mezi jednotlivci. Tyto podobnosti jsou následně projektovány do 2-D prostoru pomocí metod redukce dimenzionality, pro snaží vizualizaci vztáhů mezi jednotlivci. Následně jsou uživatelé shlukováni do dobře rozlišitelných shluků a historická data stávajících uživatel jsou použita k navzorkování vstupních dat pro nové uživatele z toho samého shluku.

Klíčová slova: t-SNE, redukce dimenzionality, podobnostní metriky nad distribucemi, nálady, detekce anomalií

1 Introduction

Knowledge of either single employees emotional state or collective mood is a key managerial task. It has been shown in earlier papers that emotions influence creativity [3], [4], decision making [1], [2] and most importantly job performance [5], [6]. Collective emotions have an impact on in-group communication [8], performance and team dynamics [7]. Usually mood state assessment is left to the manager of whom is expected to process the emotions of each individual directly. This method is not sustainable for large teams. Observing mood state of individuals in smaller teams would not be an issue in most cases, however when we take into consideration collective mood state, which is a result of a fusion of mood states of different individuals, it gets a bit tricky. To evaluate how the team is doing as a single entity requires to keep track of the mood state of each individual. Furthermore, one needs to keep in mind standard mood state of each team member and evaluate his mood fluctuations against this standard value. Standard mood state is intuitively understood as a long-term mood behaviour of an individual.

Beside actual mood state of the team members, manager is responsible also for their overall well-being, in other words how did the employee feel in the last week, last month, last year, etc. Solution for a long-term well-being assessment are supposed to be employees' self-reports of their overall affective state. These are usually taken once a year, to give the manager overall information about employees satisfaction and affective state. Unfortunately, annual employee surveys are not considered effective as they measure only long-term employees state and do not provide constant feed of the employees affective data. Therefore, both a new method based on constant feed of the employees and novel approach to employee state assessment needs to be designed.

2 Data-sets and Processing

Data-sets were gathered by the mobile and web software called EMU (Employee Mood Up), which works on a basis of employee-manager type of interactions. Prime role of the software is not individuals mood collection. Software was designed to allow effortless communication across company with focus on direct manager- employee interaction. Each employee can communicate his worries, problems or any kind of thoughts in a form of a message which is accompanied by mood state and category specifying the problem. Interactions are not restricted to messages, users are allowed to share moods or mood with category only at any time. Provided data records have been suited into a set

$$M_k = \{(m_{jk}, \tau_{jk}, c_{jk}) \in M \times T \times C | j = 1, \ldots, N_k\},$$

where N_k is number of interactions for user k.

$$M = \{m_{jk} \in \{1, 2, 3, 4, 5\} | j = 1, \ldots, N_k \wedge k = 1, \ldots, N\}$$

is a set of moods chronologically ordered by creation date for each individual. Set

$$T = \left\{\tau_{jk} \in \mathbb{R}_0^+ | j = 1, \ldots, N_k \wedge k = 1, \ldots, N\right\}$$

contains chronologically-ordered interaction creation dates. And

$$C = \{c_{jk} | j = 1, \ldots, N_k \wedge k = 1, \ldots, N\}$$

is a set of categories, specifying the mood, chronologically ordered by creation date. Creation dates are categorical variables even though time is usually seen as continuous entity. This assumption, however applies only in cases when we are considering the whole set. If we are doing for example week by week evaluations, than in terms of weeks, time is continuous variable but creation dates of each interaction within each week are seen as categorical variables with certain ordering.

Moods are discrete variables on a scale one to five. Where mood equal 1 corresponds to the worst possible state and mood with value 5 represents the best possible mood state. Participants are allowed to fill mood or message any time they wish. Beside these spontaneous interactions, participants are regularly (once a day during the weekdays) asked to fill their mood. We do not differentiate between spontaneous and induced interactions. Both are seen as current state assessments. Questionable may be the scale

for different individuals. Mood 4 for one user may not correspond to mood 4 for the other one. This would not be an issue for individual evaluations, however requirement for unified scale would be reasonable in group aggregations or user comparison. On the other hand, we assume that each individual is able to asses his current mood state on a given scale, which is common for all participants and what matters to the manager is subjective notion. If the user fills the mood 1, he personally believes, he reached his bottom and it is of no value that other individual would grade his current mood state differently, just because from his perspective, he does not perceive the situation that induced the mood as that critical. Therefore, we see interactions as subjective perceptions of ones mood state on a given scale.

For each individual we define what we call standard mood state, which is a normalised distribution of mood states for a given period of time. This quantity will be one of the key characteristics of each individual. We suspect this quantity to be time invariant. By time invariant we mean that over a long period of time the distribution will change only slightly. To reach this requirement, wide time range needs to be taken into consideration. Based on the current observations at least 2-4 months of frequent participation.

To derive such a quantity mathematically, we define distribution of transitions between mood states as 2 dimensional matrix $\mathbb{T} = (t_{ij})_{i,j=1}^{5}$, where t_{ij} is probability of occurrence of transition between states i and j. Then mood distribution is defined as sum of the column elements of transition matrix T.

3 Measuring mood state

In this section we introduce a measure for comparing mood states of two individuals. Now we start by comparing two users, by comparing their mood distributions. Commonly used measures for two distributions comparison are divergences [9], [10]. Unfortunately divergences do no take into consideration domain of the distribution. Hence, users with distributions $p = [1, 0, 0, 0, 0], q = [0, 0, 0, 0, 1]$ would in most cases have distance 0 (Jensen-Shannon [11], χ^2 [12], Hellinger divergence [13]) or infinity (Kullback-Leibler divergence [14], [15]), even though the individuals are as dissimilar as possible. Another drawback of this measures is that they do not satisfy metrics axioms.

Very convenient seems to be so called Wasserstein 1st distance also known as Earth mover's distance first defined in [16]. It has been also proven in [17] that Earth mover's distance is Mallows distance. Computing the Mallows distance is based on a solution to the transportation problem [18] and intuitively represents the shortest distance between two distributions which we will address as clusters in two-dimensional case. Let $p = \{(x_1, p_1), (x_2, p_2), \ldots, (x_m, p_m)\}$ and $q = \{(y_1, q_1), (y_2, q_2), \ldots, (y_n, q_n)\}$ be two discrete probability distributions (clusters). Let $D = (d_{i,j})_{i,j=1}^{m,n}$ be the ground distance matrix. Elements $d_{i,j}$ are called ground distances between points x_i and y_j. Typical ground distance is Euclidean distance or absolute distance. A flow between $X = [x_1, \ldots, x_m]$ and $Y = [y_1, \ldots, y_n]$ is any matrix $F = (f_{i,j})_{i,j=1}^{m,n}$. Intuitively, $f_{i,j}$ represents the amount of weight at x_i which is matched to weight at y_j. The flow F is a feasible flow between X and Y if and only if

$$f_{i,j} \geq 0 \qquad i \in [1,m], j \in [1,n] \tag{1}$$

$$\sum_{j=1}^{n} f_{i,j} \leq p_i \qquad i \in [1,m] \tag{2}$$

$$\sum_{i=1}^{m} f_{i,j} \leq q_j \qquad j \in [1,n] \tag{3}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} = \min\left(\sum_{i=1}^{m} p_i, \sum_{j=1}^{n} q_j\right). \tag{4}$$

Constraint (1) requires the amount of mass transferred from location x_i to y_j to be non-negative, which intuitively means that mass is moved from distribution p to q and not vice versa. Constraint (2) that the weight in Y matched to x_i does not exceed p_i. In other words, amount of mass supplied to Y will not exceed amount of mass in x_i. Similarly (3) limits amount of mass transferred to Y to weights q_j. Finally, constraint (4) forces the total amount of mass transferred to be equal to the weight of the lighter distribution (distribution containing less mass). The Earth mover's distance $EMD(p,q)$ between distributions p and q is the minimum amount of work to match p to q, normalized by the weight of the lighter distribution:

$$EMD(p,q) = \frac{\min_{f_{i,j}} \sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}}. \tag{5}$$

We mostly compare probability distributions (e.g. equal weight distributions), therefore constraint (4) is reduced to

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} = 1 \tag{6}$$

and the Earth Mover's distance is defined as

$$EMD(p,q) = \min_{f_{i,j}} \sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} d_{i,j}. \tag{7}$$

Beside two dimensional case, we will focus mostly on comparing discrete one-dimensional distributions with the same total weights $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$. Then the earth mover's distance is given by relation

$$EMD(p,q) = \sum_{i=1}^{j} \left| \sum_{j=1}^{n} p_j - q_j \right|, \tag{8}$$

as stated in [19]. This solution is also known as Match distance introduced first in [20], [21]. Main advantage is that the solution is now deterministic instead of solving optimization problem.

Advantage of this measure is that it works with the domain of the given distribution and furthermore it satisfies metrics axioms. It has also some convenient properties. For two distributions $p = [1, 0, 0, 0, 0], q = [0, 0, 0, 0, 1]$ is $EMD(p, q) = 4$, which is in this setting maximal possible distance. In other words individual being in the best possible state represented by distribution q (filling only the best moods) is the furthest away from the individual filling the worst mood represented by distribution p. EMD is 0 if and only if mood distributions of 2 individuals are equal. Another nice property is that EMD represents average of the distribution with domain $[4, 3, 2, 1, 0]$ if one of the compared distributions is equal $[0, 0, 0, 0, 1]$. We are however working with the distribution with domain $[1, 2, 3, 4, 5]$. Under the transformation

$$EMD_t(p, [0, 0, 0, 0, 1]) = -EMD(p, [0, 0, 0, 0, 1]) + 5,$$

$EMD_t(p, [0, 0, 0, 0, 1])$ has a meaning of an average of the distribution p with domain $[1, 2, 3, 4, 5]$, if the second distribution is given by vector $[0, 0, 0, 0, 1]$. This property will turn out to be very convenient, once we will try to categorize users. EMD metric seems like a most suitable way to measure distance between 2 individuals with respect to mood distribution. Generally any characteristics introduced in terms of probability distribution where domain is of importance can be measured by EMD metric.

4 Dissimilarity visualisations

Dissimilarity matrix also known as distance matrix is a square symmetric matrix containing pairwise distances between the elements of a set. Depending on the application, the distance used to define matrix may or may not be a metric. In case of mood distributions the measure used is EMD, which is metric.

After the distance matrix is constructed for a chosen group of individuals, visualisation of distances takes part. In particular our goal is to find an embedding in two dimensions that preserves the distances as closely as possible. If we had 15 individuals we would need visualisation in space of dimension 14 to preserve the distances exactly. Therefore, dimensionality reduction method needs to be applied to the data. One of the commonly used methods is MDS (Multidimensional scaling) [22], which is used to translate information about the pairwise distances among a set of n individuals into a configuration of n points mapped into an abstract Cartesian space. MDS is divided into 2 main types based on the used dissimilarity metric. We used precomputed version where input is directly dissimilarity matrix and no further measures are used to calculate distances between individuals. This method can be divided further into metric MDS and non-metric MDS. Second regularly used method is t-SNE [23]. Difference between these three methods is that metric MDS preserves absolute distances, non-metric version preserves ranking and t-SNE preserves local structure (controlled by parameter of perplexity, the higher the value the more uniformly distributed the points are, the lower the value, the more clustered the points are). Moreover, t-SNE has tendency to reveal data that lie in multiple, distinct clusters and reduces the tendency to crowd points together at the center. To evaluate the accuracy of the three methods, Pearson's correlation coefficient is used. Metric MDS yielded value 0.99, indicating linear relationship between real distances and

embedded ones. Non-metric MDS obtained value 0.36 and t-SNE obtained highest value of correlation for perplexity equal 7. Unfortunately no visible cluster were formed for this parameter settings. Further analysis showed that best separability of clusters and still high value of correlation is obtained for perplexity 2. Comparison of metric MDS and t-SNE with perplexity 2 is displayed in figure 1.

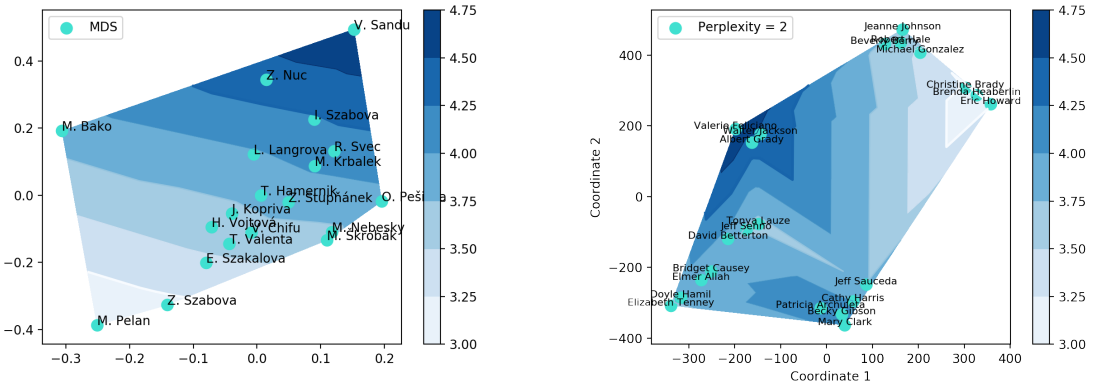


Figure 1: Visualisation of metric multidimensional scaling (figure on the left side) and t-SNE for value of perplexity equl 2 (figure on the right). Bar represents average mood.

If we were interested in detailed visualisation of similarities between users, our best choice would be metric MDS, which according to Pearson's correlation coefficient preservers distance most accurately. On the other hand, if clustering users into groups based on pairwise dissimilarity was our task, best method to use would be t-SNE which returns well distinguishable clusters.

5 Clustering

Once users form well separable clusters, we can apply suitable clustering algorithm. Generally, we have no information about the number of clusters. Furthermore, there is no training set, therefore unsupervised clustering methods need to be used. First, our intention was to cluster the very similar participants into smaller groups. However this can be done only for smaller sets containing maximum of 30 users, which corresponds either to large teams or smaller companies. For larger sets, participants tend to concentrate in the middle of the diagram, forming one compact cluster. Secondly we intended to find outliers in both smaller and larger sets containing up to 30 and over 30 participants, respectively. In small sets, individuals labeled as outliers would differ from the rest of the colleagues on team or company level. In large sets (combining data from different companies) outliers would be individuals, who differ from the rest of the population in mood.

Firstly we will focus on clustering the users into smaller groups based on their similarity. For this purpose, the clustering is applied to diagram generated by t-SNE, which returns well separable clusters. Very intuitive method to use would be k-means clustering [24], [25], algorithm dividing observations into k clusters, so each observation would

fall into cluster with the nearest mean. Downfall of this method is unknown number of clusters. Elbow method, for finding the proper number of clusters, would solve this issue. For large number of datasets, manual calculation of number of clusters, would be unsustainable. Hence, we decided to use method that does not require initial knowledge of the number of clusters. One such method is DBSCAN [26] (density based spatial clustering of applications with noise). DBSCAN is clustering algorithm grouping together observations lying close to each other (nearby neighbours) and marking as outliers points, their nearest neighbour is too distant. This method does not require knowledge of number of clusters, however it has one initial parameter r, specifying the radius of a neighborhood with respect to some point. Points are classified as outliers, if they are not situated in a reachable distance from any other point of the set. After application of t-SNE, data form clear clusters and there seem to be no obvious outliers if the dataset is of suitable size. Figure 2 illustrates application of DBSCAN on data preprocessed by t-SNE. Clustering algorithm detected 6 clusters, containing from 3 to 5 observations, each. As no specific characteristics about the participants were given, we cannot assign particular clusters any property. What the diagram says is, that users close to each other or those who find themselves in the common cluster are similar in terms of distribution of mood states.

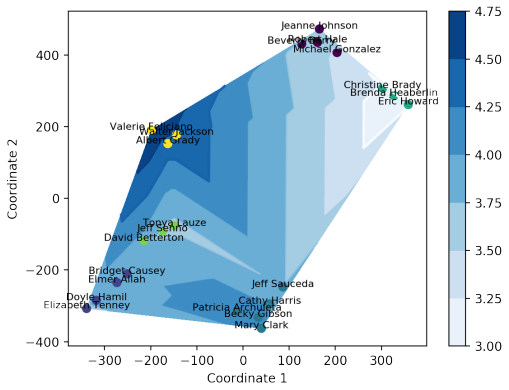


Figure 2: Application of DBSCAN method for employee clustering in combination with t-SNE method (figure on the left). Bar represents average mood.

6 Anomaly detection and cold start

Definition of the standard mood state of an individual/team allows us to define anomalies as states that differ significantly from common emotional behavior. Time series with the daily timestamp resemble random behavior, data points are too scattered. Therefore, we attempted to study anomalies with a weekly timestamp. Moods were aggregated for each individual using running average with window corresponding to 20 weeks to obtain smoother curve. In some applications of anomaly detection, average is calculated from all the historical values. Human behavior, however, is not static and tends to change over the months. Therefore we decided to take running average with window corresponding to 20 weeks instead of the whole history. Then running standard deviation was calculated

in the same manner. Anomaly detection was inspired by Nelson rules. Three rules, indicating anomaly, were defined:

1. One point (weekly average mood of an individual/team) is more than 1 standard deviations from the running average (outlier)
2. Two (or more) points (weekly average mood of an idividual/team) in a row are continually increasing (or decreasing) (trend)
3. Three (or more) points (weekly average mood of an idividual/team) in a row are on the same side of the running average (shift).

If any point or set of points satisfies on the rules, then it is labeled as anomaly. This method works fine as long as you have historical data. In most cases 3 months of observation are required. After this period of time, the standard deviation seems to level out. However, in reality you can not afford to observe an individual for 3 months and just then give some results. Added value must be immediate. Therefore, to avoid cold start, we ask each individual, at the beginning of the measurement, how he/she usually feels. We than compare this individual with our historical users, using the methods described above. Firstly we calculate similarity to other individuals using EMD metrics, then t-SNE is used for projection to 2-D space and finally DBSCAN is applied. In this way, user is assigned to one of the clusters. Individuals that share the same cluster are then taken and Gaussian KDE is generated from the empirical sample. 20 points are than sampled from this distribution and these are used as historical data for the new individual. See figure 3 for comparison of the anomaly detection with and without cold start. From the figure 3, it is obvious that including artificial historical data (black lines) enables the anomaly detection from the week 1. While the red lines, representing original data suffering from the cold start, show stable trend after first 9-12 weeks, the black lines provide stable trend from the beginning. Therefore, we do not have to wait 3 months to be able to determine standard behavior and start detecting anomalies. One may ask, why not sample the historical data from the initial distribution, the individual provides us. The reason is, that at the beginning the user needs to get used to the mood scale and when we ask him to provide his best guess how he usually feels on scale 1 to 5, it is hard to guess right. By taking similar users and sampling from their mutual distribution we bring randomness to the whole process. So even if the user is not able to guess his initial distribution correctly, we are able to cover broader range of historical observations by considering behavior of individuals similar to him.

7 Conclusion

In general, we introduced methods for employees comparison, which can be applied either to individuals or teams. Firstly basic characteristic called mood distribution, obtained from continuous long term mood measurement, are defined. Subsequently measure for mood is introduced as well as metrics for distribution comparison. Main accomplishment of this research are dissimilarity maps, as visualisation techniques for individual and team comparison with respect to different characteristics as well as method for dealing with cold start for anomaly detection.

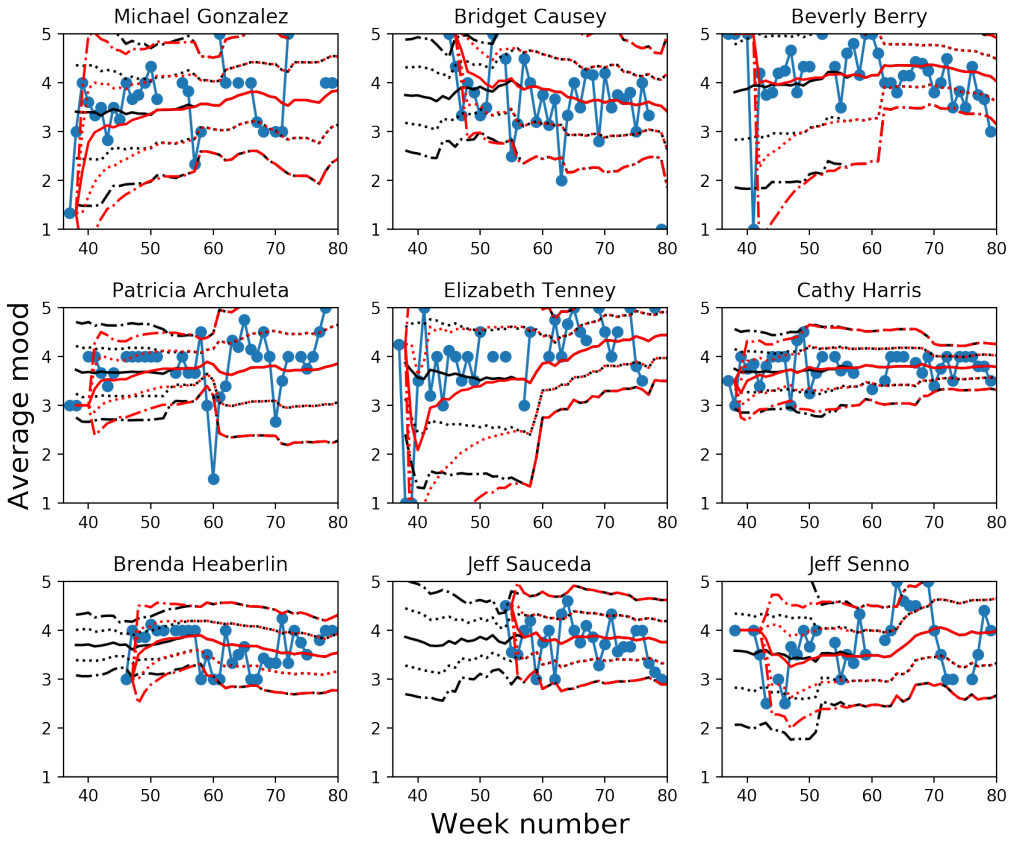


Figure 3: Figure displaying anomaly detection suffering from the cold start (red lines) and anomaly detection using artificial historical data (black lines). Blue connected dots correspond to weekly individual averages. Dotted lines and dotted-dashed lines correspond to the one and two standard deviations from the running average, respectively.

We introduce two types of maps. First one, a result of metric multi-dimensional scaling, enables to compare dissimilarity of users, while preserving absolute distances between them. However, this method does not produce clearly distinguishable clusters. Therefore, we introduce second method based on dimensionality reduction, known as t-SNE. Results of this method are highly dependable on the perplexity, as one of the initial parameters. Low values of perplexity produce well distinguishable clusters. However, values too low does not preserve absolute distances. By suitable choice of perplexity, one can achieve good differentiability and sufficient precision of absolute distances.

Further analysis has been carried out and by clustering algorithm users has been separated into groups. Two algorithms were tried out, however only one seems as a suitable choice. K-means clustering requires knowledge of the number of cluster, which we do not have. DBSCAN does not suffer from this short-comming and gives promising results.

Finally, clustering users enables us to deal with the cold start in case of anomaly

detection. New user is compared to historical users and initial data for the new user are sampled from the historical data to him similar individuals. This method enables us to detect anomalies in behavior from the very beginning.

Mood state viewed as a crucial personal characteristics does not give as an overall information about an individual. Therefore, it is desirable to introduce other characteristics (ones stability or predictability) and then study their combination. Named characteristics certainly does not cover the whole spectrum of ones personality, hence we need to define new quantities characterising each individual. Combination of different quantities could indicate same of the standardised personality characteristics as extroversion, neuroticism, oppeness to experience and so on.

But even without connection to standard personality traits these visualisations contain strong information about team members, their similarity and overall state from different points of view.

References

[1] J. P. Forgas. Mood and judgement: the affect infusion model (AIM). *Psychological Bulletin*, **117(1)**, 39, 1995.

[2] A. D. Angle, S. Connely, E.P.Waples, V.Kligyte. The influence of discrete emotions on judgement and decision-making: A meta-analytic review. *Cognition and Emotion*, **vol. 25**, 8, 2011.

[3] M. Baas, C. K. De Dreu, B.A. Nijstad. A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus?. *Psychological Bulletin*, **134(1)**, 779, 2008.

[4] F. Zenasni, T. Lubart. Effects of mood states on creativity. *Current Psychology Letters*, **8**, pp. 33-50, 2002.

[5] T. A.Wright, R. Cropanzano. The Role of Psychological Well-Being in Job Performance: A fresh look at an age-old quest. *Organizational Dynamics*, **33(4)**, pp. 338-351, 2004.

[6] R. Cropanzano, K. James, M.A. Konovsky. Dispositional afectivity as a predictor of work attitudes and job performance. *Journal of Organizational Behavior*, **14(6)**, pp. 595-606, 1993.

[7] A.P. Brief, H.M.Weiss. Organizational behavior: affect in the workplace. *Annual Review of Psychology*, **53(1)**, pp. 279-307, 2002.

[8] J. R. Kelly, S.G. Barsade. Mood and emotions in small groups and work teams. *Organizational Behavior & Human Decision Processes*, **86**, pp. 99-130, 2001.

[9] I. Csiszár. Information measure: A critical survey. *Trans. 7th Prague Conf. Inform.Theory*, Prague, 1974.

[10] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci, Math. Hungar.*, **2**, pp. 299-318, 1974.

[11] A.K.C. Wong, M.You. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, **7**, pp.599-609, 1985.

[12] H. Jeffreys. An invariant form for the prior probability in estimation problem. *Proc. Roy. Soc. Lon. Ser A.*, **186**, pp.453-461, 1946.

[13] E. Hellinger. NeueBegrundung der Theorie der quadratischen Foemen von unenlichenvielenVeranderlichen. *J. Reine Aug. Math.*, **136**, pp.210-271, 1909.

[14] S. Kullback, R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics.*, **22(1)**, pp.79-86, 1951.

[15] S. Kullback. Information Theory and Statistics. *John Willey & Sons.*, 1959: ISBN 0-8446-5625-9.

[16] Y. Rubner, C. Tomasi, L. J. Guibas. The Earth Mover's distance as a metric for image retrieval. *Technical Report STAN-CS-TN-98-86*, 1998.

[17] E. Levina, P. Bickel. The EarthMover's Distance is the Mallows Distance: Some Insights from Statistics. *Proceedings of ICCV 2001*, pp. 251-256, Vacouver, 2001.

[18] F. L. Hitchcock. The distribution of a product from several sources to numerous localities. *J. Math. Phys.*, **20**, pp. 224-230, 1941.

[19] Y. Rubner, C. Tomasi, L. J. Guibas. The Earth Mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, **40(2)**, pp. 99-121, 2000.

[20] H. C. Shen, A.K.C. Wong. Generalized texture representation and metric. *Computer, Vision, Graphics, and Image Processing*, **23**, pp. 187-206, 1983.

[21] M. Werman, S. Peleg, A. Rosenfeld. A distance metric for multi-dimensional histograms. *Computer, Vision, Graphics, and Image Processing*, **32**, pp. 328-336, 1985.

[22] A. Mead. Review of the Development of Multidimensional Scaling Methods. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **41(1)**, pp. 27-39, 1992.

[23] L.J.P. van der Maaten, G.E. Hinton. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, **9**, pp. 2579-2605, 2008.

[24] S.P.Loyd. Least square quantization in PCM. *IEEE Transactions on Information Theory*, **28(2)**, pp. 129-137, 1982.

[25] E.W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, **21(3)**, pp. 768-769, 1965.

[26] M. Ester, H.P.Kriegel, J. Sander, X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226-231, 1996.

T_1 Estimation Method Based on Bloch Equations Simulation

Kateřina Škardová*

2nd year of PGS, email: katerina.skardova@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

Tomáš Pevný, Faculty of Electrical Engineering, CTU in Prague

Abstract. In this work we propose an approach to T_1 relaxation time estimation, that incorporates the Bloch equations. The Bloch equations, first published in [2], describe the evolution of magnetization vector $\vec{M} = (M_x, M_y, M_z)$ in time, depending on T_1 and T_2 relaxation times. The equations can be used to simulate the signal generated in magnetic resonance. In the proposed approach we try to solve the inverse problem and estimate T_1 relaxation time from the knowledge of signal generated by vector $\vec{M}$ in eleven given times.

Keywords: Bloch equations, T_1 relaxation time, optimisation, neural networks

Abstrakt. Tato práce navrhuje metodu pro odhad T_1 relaxačního času s vyžitím Blochových rovnic. Blochovy rovnice byli poprvé představeny v článku [2] a popisují vývoj vektoru magnetizace $\vec{M} = (M_x, M_y, M_z)$, v závislosti na T_1 a T_2 relaxačních časech. Tyto rovnice mohou sloužit ke simulování signálu generovaného v magnetické rezonanci. Navrhovaný postup je založen na řešení inverzního problému, kdy je hodnota T_1 času odhadována na základě signálu generovaného vektorem magnetizace $\vec{M}$ v jedenácti časech.

Klíčová slova: Blochovy rovnice, T_1 relaxační čas, optimalizace, neuronové sítě

1 Introduction

Magnetic resonance imaging (MRI) is a frequently used modality in medical imaging. There are several advantages of MRI, compared to other major imaging modalities such as X-ray, computerized tomography (CT), positron emission tomography (PET) and ultrasound [6]. Firstly, due to the absence of ionizing radiation, MRI is a non-invasive imaging method. Secondly, the MRI is not limited by problems such as finite penetration depth or internal reflection. It also provides good soft-tissue contrast, that can be adjusted by modifying the pattern of excitation pulses. On the other hand, there are also some disadvantages of MRI, the main one being the low acquisition speed.

Myocardial T_1 relaxation time is a representative marker for a number of pathological conditions [7]. It is currently used in quantification of myocardial fibrosis.

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/194/OHK4/3T/14 and by grant MZ NV19-08-00071.

Many imaging techniques were developed for T_1 mapping. One of the most frequently used is Modified Look-Locker Imaging (MOLLI) [6] sequence, which allows for T_1 mapping of the myocardium with relatively high spatial resolution within a single breath-hold. The general principle of myocardial T_1 mapping is acquiring several images in different times after excitation pulse and performing pixel-wise fit of the image intensity with the theoretically derived equation for T_1 relaxation:

$$f(t) = A - Be^{-\frac{t}{T_1^*}}, \tag{1}$$

where A, B are fitting parameters. The "apparent" T_1^* value obtained by the fit is obviously dependent on the chosen type of sequence. In the case of MOLLI sequence, there is systematic undervaluation, which is corrected by the following formula:

$$T_1 \approx \left(\frac{B}{A} - 1\right) T_1^*. \tag{2}$$

This correction is simple and widely used, although it is known not to be theoretically justified [3], because it was originally derived for different type of sequence.

There are also more precise methods, such as Inversion Recovery (IR) method [3], which is often considered to be the "gold standard". However acquisition using IR method is too slow to be used on patients.

The goal of this work is to propose an estimation method that would eliminate this systematic error and provide values closer to the real T_1 relaxation time.

In next section, the physical meaning of T_1 and T_2 relaxation times is described and key principles of MRI are pointed out. In the following section Bloch equations are introduced and solution for the case of static magnetic field is derived. Then, MOLLI sequence and the way it was implemented is described in detail. In the last two sections the new estimation method is proposed and preliminary results are provided.

2 T_1 and T_2 relaxation

In MRI, the studied object is placed into a magnetic field with magnetization $\vec{B}_0$. This induces a macroscopic magnetization $\vec{M}_0$ in the object, which is denoted as equilibrium magnetization. Vector $\vec{M}_0$ has the same direction as the magnetic field $\vec{B}_0$, which is called the longitudinal direction. The longitudinal direction is generally denoted by the z-axis. This notation is used also in this work.

For given magnetisation $\vec{M}$, only the projection to x-y plane can be measured in magnetic resonance. This plane is denoted as transverse. To create measurable $\vec{M}$, radio-frequency (RF) pulse which deflects the magnetization M_0 from the longitudinal direction, can be applied. When the RF pulse is switched off, the magnetization vector returns to its equilibrium position. The process of z-component returning to its maximal value is called T_1 or longitudinal relaxation. The time needed for z-component to recover 63% of it maximal value is denoted as T_1 relaxation time. The simultaneous decay of the transverse component is called T_2 relaxation time. The time needed for transverse component to decay to 37% of its maximal size is denoted as T_2 relaxation time. All named stages can be seen in Figure 1.

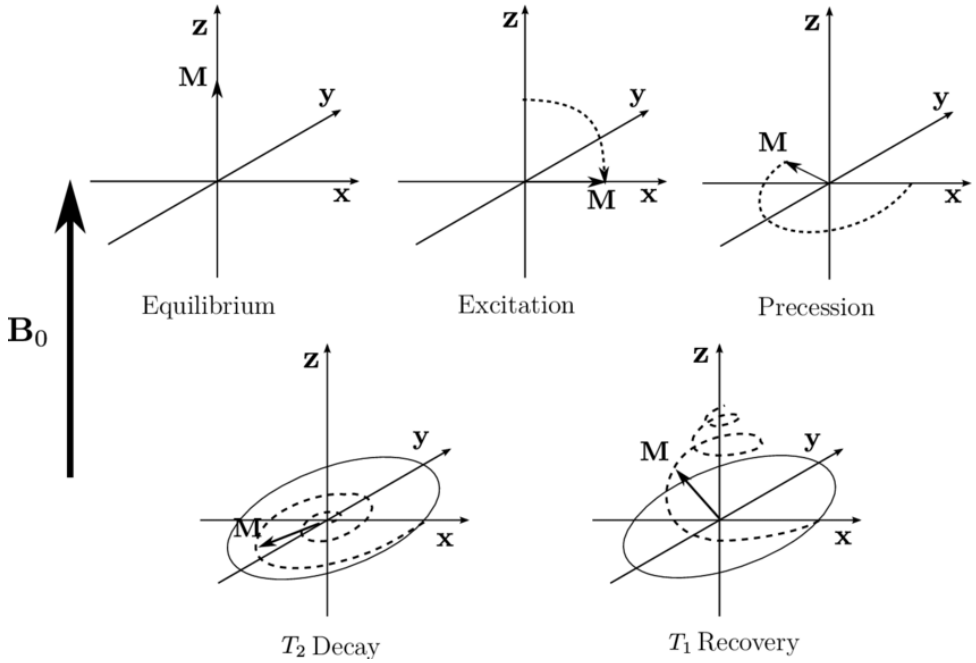


Figure 1: First line: the diagrams of $\vec{M}$ in equilibrium position, deflected by the RF pulse and precessing back to equilibrium position. Second line: the transverse and longitudinal relaxation. Reprinted from [4].

3 Bloch equations

For time dependent magnetization $\vec{M}(t) = (M_x(t), M_y(t), M_z(t))$, the Bloch equations [2] are formulated:

$$\frac{\mathrm{d}M_x(t)}{\mathrm{dt}} = \gamma(M(t) \times B(t))_x - \frac{M_x(t)}{T_2}, \tag{3}$$

$$\frac{\mathrm{d}M_y(t)}{\mathrm{dt}} = \gamma(M(t) \times B(t))_y - \frac{M_y(t)}{T_2}, \tag{4}$$

$$\frac{\mathrm{d}M_x(t)}{\mathrm{dt}} = \gamma(M(t) \times B(t))_z - \frac{M_z(t) - M_0}{T_1}, \tag{5}$$

where γ is constant characteristic of a given chemical element, $\vec{B}(t)$ is the magnetization of the external field and M_0 is the z-component of equilibrium magnetization $\vec{M}_0 = (0, 0, M_0)$.

Let us consider the case of static magnetic field: $B(t) = (0, 0, B_0),\ \forall t$. Under such condition, the solution can be found in several steps. First, the equations are simplified:

$$\frac{\mathrm{d}M_x(t)}{\mathrm{dt}} = \gamma M_y(t) B_0 - \frac{M_x(t)}{T_2}, \tag{6}$$

$$\frac{\mathrm{d}M_y(t)}{\mathrm{dt}} = -\gamma M_x(t) B_0 - \frac{M_y(t)}{T_2}, \tag{7}$$

$$\frac{\mathrm{d}M_z(t)}{\mathrm{dt}} = -\frac{M_z(t) - M_0}{T_1}. \tag{8}$$

The factor γB_0 is called Larmor frequency and is denoted by ω_0.
Using notation $M_{xy} = M_x + iM_y$, equations (6) and (7) can be combined into equation:

$$\frac{\mathrm{d}M_xy(t)}{\mathrm{dt}} = \omega_0 M_y(t) - \frac{M_x(t)}{T_2} + i\left(-\omega_0 M_x - \frac{M_y(t)}{T_2}\right) = M_{xy}\left(-i\omega_0 - \frac{1}{T_2}\right). \quad (9)$$

Ordinary differential equation (9) has solution:

$$M_{xy}(t) = M_{xy}(0)\mathrm{e}^{\left(-i\omega_0 - \frac{1}{T_2}\right)t} = \mathrm{e}^{-\frac{t}{T_2}}(M_x(0) + iM_y(0))(\cos(\omega_0 t) - i\sin(\omega_0 t)). \quad (10)$$

from which M_x and M_y can be expressed:

$$M_x(t) = \mathrm{e}^{-\frac{t}{T_2}}(M_x(0)\cos(\omega_0 t) + M_y(0)\sin(\omega_0 t)), \quad (11)$$

$$M_y(t) = \mathrm{e}^{-\frac{t}{T_2}}(M_y(0)\cos(\omega_0 t) - M_x(0)\sin(\omega_0 t)). \quad (12)$$

Since M_0 is constant, following equation can be derived from (8):

$$\frac{\mathrm{d}(M_z(t) - M_0)}{\mathrm{dt}} = -\frac{1}{T_1}(M_z(t) - M_0). \quad (13)$$

The solution of equation (13) is in following form:

$$M_z(t) - M_0 = (M_z(0) - M_0)\mathrm{e}^{-\frac{t}{T_1}}, \quad (14)$$

$$M_z(t) = M_0\left(1 - \mathrm{e}^{-\frac{t}{T_1}}\right) + M_z(0)\mathrm{e}^{-\frac{t}{T_1}}. \quad (15)$$

By combining equations (11), (12) and (15) the following set of equations is obtained:

$$M_x(t) = \mathrm{e}^{-\frac{t}{T_2}}(M_x(0)\cos(\omega_0 t) + M_y(0)\sin(\omega_0 t)), \quad (16)$$

$$M_y(t) = \mathrm{e}^{-\frac{t}{T_2}}(M_y(0)\cos(\omega_0 t) - M_x(0)\sin(\omega_0 t)), \quad (17)$$

$$M_z(t) = M_z(0)\mathrm{e}^{-\frac{t}{T_1}} + M_0\left(1 - \mathrm{e}^{-\frac{t}{T_1}}\right). \quad (18)$$

The set of equations can be also formulated in a matrix form:

$$\vec{M}(t) = \mathbb{A}(t)_{relax}\mathbb{R}_z(t)\vec{M}(0) + \vec{C}(t), \quad (19)$$

where

$$\mathbb{A}_{relax} = \begin{pmatrix} \mathrm{e}^{-\frac{t}{T_2}} & 0 & 0 \\ 0 & \mathrm{e}^{-\frac{t}{T_2}} & 0 \\ 0 & 0 & \mathrm{e}^{-\frac{t}{T_1}} \end{pmatrix}, \mathbb{R}_z = \begin{pmatrix} \cos(\omega_0 t) & \sin(\omega_0 t) & 0 \\ -\sin(\omega_0 t) & \cos(\omega_0 t) & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\vec{C} = \begin{pmatrix} 0 \\ 0 \\ M_0\left(1 - \mathrm{e}^{-\frac{t}{T_1}}\right) \end{pmatrix}.$$

Following useful equation can also be derived:

$$\vec{M}(t + \Delta t) = \mathbb{A}(\Delta t)_{relax}\mathbb{R}_z(\Delta t)\vec{M}(t) + \vec{C}(\Delta t), \quad (20)$$

Matrix $\mathbb{A}_{relax}$ represents the T_1 relaxation of z-component and T_2 relaxation of transverse components, respectively. Matrix $\mathbb{R}_z$ represents the procession around z-axis during the relaxation. Additive term $\vec{C}$ further modifies the relaxation of longitudinal component.

4 MOLLI sequence implementation

Before the sequence starts, the magnetization vector is aligned with the outer magnetic field, which is considered to be static. Therefore $\vec{M}(0) = (0, 0, M_0)$. The MOLLI sequence consists of 11 images that are acquired in three sets of 3, 3 and 5 images or 5, 3 and 3 images. Each sub sequence is started with inversion pulse, which rotates $\vec{M}$ by 180 degrees around x-axis. The angle by which $\vec{M}$ ir rotated around x-axis is called flip angle. Rotation by flip angle α can be written in matrix form:

$$\vec{M}(t) = \mathbb{R}_x \vec{M}(0), \quad \text{where } \mathbb{R}_x = \begin{pmatrix} 0 & 0 & 1 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix}, \tag{21}$$

In the subset, each image is acquired in one heartbeat and each subset is followed by 3-heartbeats pause. In 5-3-3 MOLLI sequence, the images in first subset are acquired in times $d_1, d_1 + RR, \ldots, d_1 + 4RR$, where d_1 is the delay after the inversion pulse and RR denotes the length of cardiac cycle. Second subset is acquired in times $d_2, d_2 + RR, d_2 + 2RR$ after the second inversion pulse and the third one in times $d_3, d_3 + RR, d_3 + 2RR$ after third inversion pulse. The delay times satisfy condition $d_1 < d_2 < d_3 < RR$. This way 11 different times after inverse pulse are sampled. For the final MOLLI sequence, the images are re-ordered by the time from closest inversion pulse. The scheme of 3-3-5 MOLLI sequence and final re-ordering of images can be seen in Figure 2.

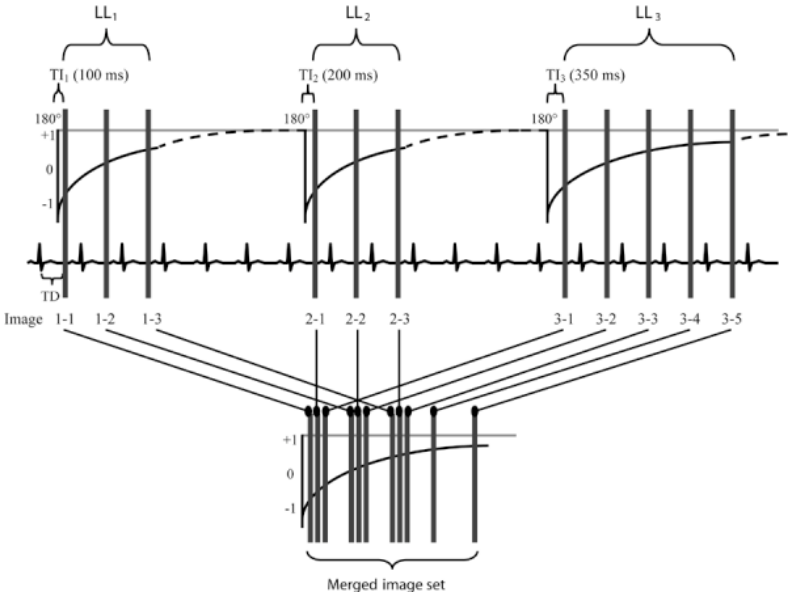


Figure 2: The scheme of 3-3-5 MOLLI sequence. In this case the delay times are $d_1 =$ 100ms, $d_2 = 200$ms and $d_3 = 350$ms. The images are acquired in three sub sequences and then re-ordered by their time after inversion pulse. Reprinted from [5].

The RR of all cardiac cycles is not the same in reality, and therefore knowing single RR and $d_1,\ d_2,\ d_3$ is not sufficient to reconstruct MOLLI sequence. In medical practice all 11 inversion times (TI) as well as the corresponding cardiac cycle durations are measured

and recorded. However, the durations of three-heart-beats pauses are not measured. In this work we use RR obtained by averaging measured cardiac cycles.

The process of image acquisition consists of two steps. First one is a preparation sequence of five RF pulses with flip angles $\frac{\alpha}{6}, -\frac{\alpha}{3}, \frac{\alpha}{2}, -\frac{2\alpha}{3}, \frac{5\alpha}{6}$. In the second step, image is aquired during balanced steady-state free precession (bSSFP). In this work bSSFP consists of 40 RF pulses with flip angle $-\alpha$ and α alternatively. The signal equal to $\sqrt{(M_x^2 + M_x^2)}$ is acquired after the fifth pulse.

In MRI, each pixel of the image is acquired from a volume of the imaged object. The dimensions of this volume depend on the image resolution. In most cases, the z-dimension of imaged volume is larger than the other two dimensions. Therefore, the flip angle α can not be considered constant through z-dimension of imaged volume. In order to simulate the real MOLLI sequence more accurately, the imaged volume was divided to ten layers with different flip angles.

For given theoretical flip angle α and layers $l_i \in \{-4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$ the flip angle in each layer is given by

$$\tilde{\alpha}_i = \alpha e^{-\frac{l_i}{2\sigma^2}}, \quad \text{for } i = 1, .., 10. \tag{22}$$

The final signal is generated as average of signals from these ten layers.

All steps of 5-3-3 MOLLI sequence in a simplified form can be seen in Algorithm 1.

Function `Bloch_MOLLI`(M_0, T_1, T_2, α, RR, ω_0, $\{l_n\}_{n=1}^{11}$, σ^2, $\{TI_n\}_{n=1}^{11}$),:
- $\{\tilde{\alpha}_n\}_{n=1}^{11} \leftarrow$ generate_flip_angle($\{\tilde{l}_n\}_{n=1}^{11}, \sigma^2$) `// computed from eq.(22)`
- $t \leftarrow 0$,
- $\vec{M} \leftarrow (0, 0, M_0)$
- type = [5,3,3] `// 5-3-3 MOLLI seq. simulated`
- **for** *subset = 1:3* **do**
  - $\vec{M} \leftarrow$ inversion pulse($\vec{M}$) `// computed from eq.(21)`
  - **for** *image = 1: type[subset]* **do**
    - $\Delta t \leftarrow$time to next preparation sequence ($\{TI_n\}_{n=1}^{11}$)
    - $\vec{M}(t + \Delta t) \leftarrow \mathbb{A}(\Delta t)_{relax}\mathbb{R}_z(\Delta t)\vec{M}(t) + \vec{C}(\Delta t)$
    - $\vec{M} \leftarrow$ preparation sequence ($\vec{M}, \{\frac{\alpha}{6}, -\frac{\alpha}{3}, \frac{\alpha}{2}, -\frac{2\alpha}{3}, \frac{5\alpha}{6}\}$)
    - **for** *i = 1:10* **do**
      - $\vec{M} \leftarrow$ bSSFP sequence($\vec{M}, \tilde{\alpha}_i$)
      - layers[] $\leftarrow \sqrt{(M_x^2 + M_x^2)}$
    - **end**
    - signal[] $\leftarrow$ average(layers[]) `// averaging signal from 10 layers`
  - **end**
  - $\Delta t \leftarrow 3RR$ `// 3-heart-beat pause`
  - $\vec{M}(t + \Delta t) \leftarrow \mathbb{A}(\Delta t)_{relax}\mathbb{R}_z(\Delta t)\vec{M}(t) + \vec{C}(\Delta t)$
- **end**
- signal[] $\leftarrow$ re-order(signals[]) `// re-order by inversion time`
- **return** signal[]

End Function

Algorithm 1: Algorithm for simulation of 5-3-3 MOLLI sequence.

5 Proposed method

The proposed method for T_1 relaxation time estimation is obtained by neural network trained of simulated data.

MOLLI sequence simulator described in previous section, is used for artificial data generation. As can be seen in Algorithm 1, the simulator has following arguments: M_0 (equilibrium magnetization), T_1 (relaxation time), T_2 (relaxation time), α (flip angle), RR (length of cardiac cycle), ω_0 (precession frequency), σ^2 (flip angle variance), $\{l_n\}_{n=1}^{11}$ (imaged layers) and $\{TI_n\}_{n=1}^{11}$ (set of inversion times). In order to generate training data set, all parameters were randomly generated in ranges occurring in real MOLLI sequences. Specifically, values of T_1 in range 300 - 1800, T_2 in range 50 - 400, M_0 in range 0.1 - 0.95, RR in range 700-1200. The inversion times we generated based on RR values and delay times $d_1 = 100,\ d_2 = 180,\ d_3 = 260$, because these values were used in all our real data sets. For the same reason the flip angle was set to 35 degrees in all simulations.

Neural network with six dense layers and ReLu activation function [1] was used to provide results presented in this work. The neural network was trained on batches of 150 sets of values generated by the MOLLI simulator.

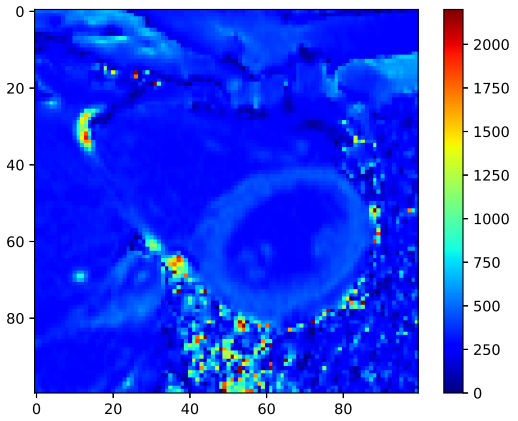
6 Experimental results

At this point only preliminary results were obtained by the proposed method. The T_1 relaxation times estimated by the proposed method are compared with values generated by magnetic resonance based on the direct fitting algorithm described in first section of this work. Two norms were used to measure the difference between MR generated data ($\vec{x}_i$) and estimated data ($\vec{y}_i$). Normalised difference: $(\sum_{i=1}^{n}((x_i - x_i)/y_i)^2)^{1/2}$ and absolute difference $(\sum_{i=1}^{n}((x_i - x_i))^2)^{1/2}$. The mean and absolute value (per pixel) is also provided in the following results.

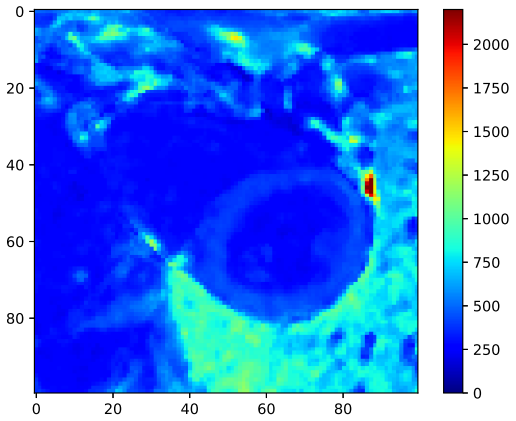
The comparison of MR generated T_1 relaxation times and T_1 times obtained by the first step of proposed method can be seen in Figure 3. In myocardium area, the estimation is similar to values generated by MR. Different values were estimated in the background, where the proposed method provided significantly higher values. In Figure 3d, the estimated values are plotted as a function of values generated in magnetic resonance.

7 Conclusion

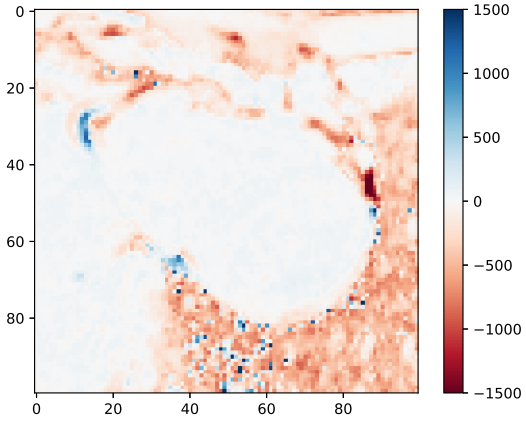
The goal of this work was to propose a method that would provide T_1 relaxation time estimation without the systematic error that is currently present in values generated by magnetic resonance. In the proposed method, the estimation is provided by neural network trained on artificial data. For artificial data generation, MOLLI sequence simulator incorporation Bloch equations was implemented. The results of proposed method were compared with the T_1 values generated by magnetic resonance. In the myocardium, similar T_1 relaxation times were obtained. In general, higher values were generated by the proposed method. This observation is in agreement with the fact that MR generated



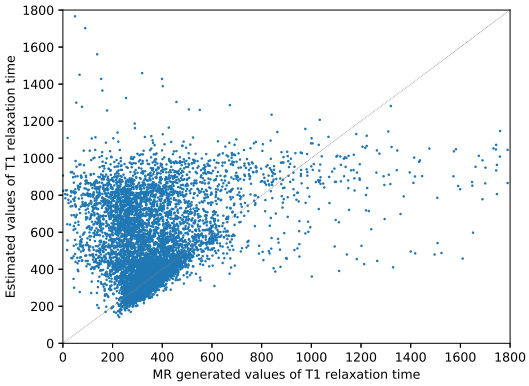
(a) Map of T_1 times produced by MR.



(b) Map of estimated T_1 times.



(c) Difference between MR and estimated T_1 times.



(d) The relation between MR-generated and estimated T_1 relaxation times.

Figure 3: Comparison of T_1 map generated by MR and map generated by the proposed method. The normalised difference between generated and estimated values = $5.2 \cdot 10^9$, mean absolute difference = 322.81 , median absolute difference = 141.82.

values are known to be systematically undervalued. Testing the proposed method on more real data sets is needed in the future work. Adding a second step to the method might also be beneficial to improve the final estimation.

References

[1] C. C. Aggarwal, "Neural networks and deep learning," *Cham: Springer International Publishing*, 2018.

[2] F. Bloch, "Nuclear induction," *Physical review*, vol. 70, no. 7-8, p. 460, 1946.

[3] M. A. Cooper, T. D. Nguyen, P. Spincemaille, M. R. Prince, J. W. Weinsaft, and Y. Wang, "How accurate is molli t1 mapping in vivo? validation by spin echo methods," *PloS one*, vol. 9, no. 9, p. e107327, 2014.

[4] A. Hazra, G. Lube, and H.-G. Raumer, "Numerical simulation of bloch equations for dynamic magnetic resonance imaging," *Applied Numerical Mathematics*, vol. 123, pp. 241–255, 2018.

[5] D. R. Messroghli, A. Radjenovic, S. Kozerke, D. M. Higgins, M. U. Sivananthan, and J. P. Ridgway, "Modified look-locker inversion recovery (molli) for high-resolution t1 mapping of the heart," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 52, no. 1, pp. 141–146, 2004.

[6] D. G. Nishimura, *Principles of magnetic resonance imaging*. Stanford University, 1996.

[7] A. J. Taylor, M. Salerno, R. Dharmakumar, and M. Jerosch-Herold, "T1 mapping: basic techniques and clinical applications," *JACC: Cardiovascular Imaging*, vol. 9, no. 1, pp. 67–81, 2016.

Detection of Alfvén Eigenmodes on the COMPASS Tokamak with Neural Networks*

Vít Škvára

4th 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. Tokamak is an experimental device for research of thermonuclear fusion. High temperature and energy plasma is confined in the tokamak chamber by strong magnetic forces. One of the leading issues of tokamak fusion is the instability of the plasma column, where an uncontrolled disruption may possibly endanger the plasma facing components. There are many types of instabilities with Alfvén eigenmodes being the one of our interest. Alfvén eigenmodes can be detected from spectrograms of certain magnetic probes. In our work we train a two stage model on both labeled and unlabeled patches of spectrograms. The first stage is a generative neural network based on the convolutional Variational Autoencoder that produces a low dimensional latent representation of the input data. Multiple ways of shaping the latent space via choices of prior distributions and different loss functions are explored in order to improve the performance of the second stage model, which is a classifier trained with labeled data. On a number of experiments, we show that our approach is a viable option for automated detection of rare instabilities in tokamak plasma.

Keywords: anomaly detection, generative models, neural networks, tokamak fusion

Abstrakt. Tokamak je experimentální zařízení pro výzkum termojaderné fúze. Vysokoteplotní plazma je při experimentu udržováno v komoře tokamaku silnými magnetickými silami. Jednou z nějvětších překážek při experimentálním provozu je vysoká nestabilita plazmatu. Při nekontrolované ztrátě udržení může dojít i k poškození vnitřku komory. Existuje mnoho příčin a typů nestabilit, z nichž nejzajímavější je tzv. Alfvénovský mód. Ten lze pozorovat na spektrogramu specifických magnetických sond. V této práci představujeme dvoustupňový model trénovaný na olabelovaných i neolabelovaných spektrogramech. První stupeň je generativní neuronová síť na principu konvolučního variačního autoencoderu, která vstupní data promítá do latentního prostoru snížené dimenze. V práci je popsáno několik různých způsobů jak přizpůsobit tvar výsledného latentního prostoru pomocí použití různých aprioren a ztrátových funkcí tak, aby byla usnadněna úloha sekundární klasifikátoru, trénovaného s olabelovanými daty. Na několika experimentech je demonstrována schopnost celého algoritmu úspěšně detekovat vzácné jevy v experimentálních datech.

Klíčová slova: detekce anomálií, generativní modely, neuronové sítě, termojaderná fúze

Full paper: V. Škvára, T. Pevný, V. Šmídl, J. Seidl, A. Havránek. *Detection of Alfvén eigenmodes on COMPASS with neural networks.* Submitted to Fusion Science and Technology (2019).

*This work has been supported by the GACR project 18-21409S

Deep Ensemble Filter for Active Learning

Lukáš Ulrych

3rd year of PGS, email: ulrycluk@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. The design of existing techniques for active learning does not take into account the incremental nature of the task. Ensemble filters on the other hand do, by utilizing Bayes' rule, but they are more concerned with close approximation of the posterior distribution and do not offer good estimation of variance which is needed for state space exploration. Based on Kalman filter, we propose Deep ensemble filter (DEnFi). Key idea of DEnFi is to evolve an ensemble of neural networks by iterating two steps: inflation and localization. On examples from Bayesian optimization and Active classification we show the superiority of DEnFi in regards to finding the true minimum and to provide good classification with correct uncertainty estimation, respectively.

Keywords: Active learning, Ensemble filter, Deep neural networks

Abstrakt. V současnosti využívané techniky aktivního učení nijak nevyužívají inkrementální podstatu jeho problému. Ansámbl filtry toho jsou schopny díky Bayesově větě, bývají ale zaměřeny spíše na co nejlepší aproximaci módů aposteriorní distribuce a nehodí se k odhadům variance, které jsou potřeba k dobrému prohledávání daného prostoru. Na základech Kalmanova filtru navrhujeme Deep ensemble filter (DEnFi). Klíčová myšlenka našeho přístupu je založena na vývoji ansámblu tvořeného neuronovými sítěmi iterováním dvou kroků: inflace a lokalizace. Na příkladu Bayesovské optimalizace a Aktivní klasifikace ukážeme výsledky, jakých je DEnFi schopno dosáhnout v porovnání s jinými metodami.

Klíčová slova: Aktivní učení, Ansámbl filtrace, Hluboké neuronové sítě

Full paper: L. Ulrych, V. Šmídl. *Deep ensemble filter for active learning.* Submitted to 'Bayesian Deep Learning Workshop 2019'.

Ruling Principles for Decision-Based Pedestrian Model*

Jana Vacková

3rd year of PGS, email: janca.vackova@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

Marek Bukáček, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. To predict the pedestrian movement during evacuations or any other kind of crowd situations, the pedestrian models are used. Individual data analysis from egress experiments stands beside the modeling, but it is necessary as well. These two elements, models and data analysis, form the principal goals of the research in this field. The global aim of our research is to connect these two elements - to use conclusions obtained from the microscopic data analysis [7, 8] about following in the crowd or [2] about individual strategies to make the models more accurate. Enhanced rules at all decision-making levels make the model more realistic from a microscopic perspective, thus its predictive power increases.

From the motivation mentioned above, we build continuous (in the both time and space) stochastic model to maintain the real human behaviour pattern using the microscopic point of view [4] due to the possibility to implement the individual phenomena in the future. To capture the decision-making process properly [1], we let the pedestrians make their own decisions in accordance with conditions in their neighbourhood. It means that the rules are applied to every single pedestrian separately in a specific (deterministic, random or more sophisticated) order. This kind of rules results in the microscopic solution of pedestrian collisions even in the dense crowd (when arches can occur [3]) also in a microscopic way, i.e. without moving any other pedestrian than the affected ones right in the solving situation.

Besides the package of movement rules, we present our calibration scheme for finding values of model parameters that produce real behaviour - we perform a calibration episode which proves the possibility to calibrate the model in real macroscopic values [5, 6]. The complete calibration process using data from E4 [2] is in progress and will be discussed in the future work.

Keywords: Pedestrian dynamics, Decision-based model, Calibration

Abstrakt. Pro predikci pohybu chodců nejen při evakuacích se používají chodecké modely. Stejně tak důležitá jako samotné modelování je datová analýza jednotlivců z evakuačních experimentů. Tyto dva elementy, modelování pohybu a analýza dat z experimentů, formují hlavní cíle výzkumu v této oblasti. Globální cíl našeho výzkumu je propojení těchto dvou elementů, a to použití závěrů získaných z mikroskopické analýzy dat [7, 8] o fenoménu následování v davu či [2] o individuálních strategií k zpřesnění modelů. Vylepšená pravidla v rámci všech úrovní

*This work has been supported by the Grant SGS18/188/OHK4/3T/14 provided by the Ministry of Education, Youth, and Sports of the Czech Republic (MŠMT ČR).

rozhodovacího procesu chodce učiní model realističtější z mikroskopické perspektivy, což zvýší jeho predikční sílu.

Dle výše popsané motivace budujeme stochastický model spojitý jak v čase, tak v prostoru kvůli zachování vzorce reálného lidského chování na mikroskopické úrovni [4] a možnosti implementovat do něj v budoucnu různé individuální fenomény. Abychom detailně zachytili rozhodovací proces [1], necháváme rozhodovat chodce v souladu s jejich okolními podmínkami. To znamená, že pravidla aplikujeme zvlášť na každého jednotlivého chodce ve specifickém (deterministickém, náhodném či jiném) pořadí. Tato pravidla vedou na řešení kolizí chodců na mikroskopické úrovni dokonce i v hustém davu (za přítomnosti můstku [3]), tj. bez hýbání jiného chodce než toho, kterého se týká právě řešená situace.

Vedle této sady pravidel představujeme užívané kalibrační schéma pro nalezení hodnot parametrů modelu, které produkuje jeho reálné chování - popisujeme kalibrační epizodu, která prokazuje možnost kalibrovat model na reálné makroskopické hodnoty [5, 6]. Kompletní kalibrační proces využívající data z E4 [2] momentálně probíhá a bude obsahem budoucí práce.

Klíčová slova: Pohyb chodců, Pravidlový model, Kalibrace

Full paper:

- Vacková, Jana and Bukáček, Marek. Ruling Principles for Decision-Based Pedestrian Model. *Stochastic and Physical Monitoring Systems*, 2019.
- Vacková, Jana and Bukáček, Marek. The Microscopic Analysis of Velocity-Density Paradigm. *International Conference on Applied Mathematics*, 2019.
- Vacková, Jana and Bukáček, Marek. Follower-Leader Concept in Microscopic Analysis of Pedestrian Movement in a Crowd. *Pedestrian and Evacuation Dynamics 2018*, Springer, 2019.

References

[1] Bailo, Rafael and Carrillo, José A and Degond, Pierre. Pedestrian Models Based on Rational Behaviour. *Crowd Dynamics, Volume 1*, 259–292, Springer, 2018.

[2] Bukáček, Marek and Hrabák, Pavel and Krbálek, Milan. Microscopic travel-time analysis of bottleneck experiments. *Transportmetrica A: transport science*, **14**, 375–391, Taylor & Francis, 2018.

[3] Garcimartín, A and Zuriguel, I and Pastor, JM and Martín-Gómez, C and Parisi, DR. Experimental evidence of the "Faster Is Slower" effect. *Transportation Research Procedia*, 760–767, Elsevier, 2014.

[4] Kang, Wonho and Han, Youngnam. A Simple and Realistic Pedestrian Model for Crowd Simulation and Application. *arXiv preprint arXiv:1708.03080*, 2017.

[5] Schadschneider, Andreas and Chraibi, Mohcine and Seyfried, Armin and Tordeux, Antoine and Zhang, Jun. Pedestrian Dynamics: From Empirical Results to Modeling. *Crowd Dynamics, Volume 1*, 63–102, Springer, 2018.

[6] Sparnaaij, Martijn and Duives, Dorine C and Knoop, Victor L and Hoogendoorn, Serge P. Multiobjective Calibration Framework for Pedestrian Simulation Models: A study on the Effect of Movement Base Cases, Metrics, and Density Levels. *Journal of Advanced Transportation*, Hindawi, 2019.

[7] Vacková, Jana and Bukáček, Marek. Follower-Leader Concept in Microscopic Analysis of Pedestrian Movement in a Crowd. *Pedestrian and Evacuation Dynamics 2018*, Springer, 2019.

[8] Vacková, Jana and Bukáček, Marek. The Microscopic Analysis of Velocity-Density Paradigm. *International Conference on Applied Mathematics*, 2019.

[9] Vacková, Jana and Bukáček, Marek. Ruling Principles for Decision-Based Pedestrian Model. *Stochastic and Physical Monitoring Systems*, 2019.

3D simulace růstu krystalů s užitím rovnice fázového pole*

Aleš Wodecki

1. ročník PGS, email: aleswodecki@gmail.com
Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitelé:

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

Pavel Strachota, Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Abstract. We deal with numerical solution of a three-dimensional phase field model of solidification in single component anisotropic material. In this contribution, we extend the model by a crystal orientation transformation. This transformation allows us to simulate the growth of dozens of crystals with different orientations inside of a domain. The transformation is governed by an orientation field that dynamically changes based on the phase field within the domain. Finsler geometry is used to simulate anisotropic crystal growth with good results. A robust algorithm is developed to simulate the growth of multiple grains with an arbitrary number of random crystallographic orientations and a fully resolved 3D dendritic geometry. This algorithm uses MPI and openMP to provide parallelisation across threads and nodes with decent scaling performance and a two dimensional division of the three dimensional domain. In the first part, the model and the parallel implementation of the algorithms are explained. The second part is devoted to demonstrating the effect of mesh-related numerical anisotropy and the simulations of complex polycrystalline solidification on very fine meshes. All of the simulations are performed using the finite volume method on a regular cuboid mesh.

Keywords: phase field, finite volume method, anisotropic crystal growth

Abstrakt. Třídimensionální rovnice fázového pole lze užít k modelování tuhnutí jednosložkového anisotropního materiálu. V tomto příspěvku rozšiřujeme náš model o transformaci krystalové orientace. Tato transformace nám umožňuje simulaci růstu desítek krystalů o různých orientacích v jedné oblasti. Tato transformace je řízena orientačním polem, které se dynamicky mění v závislosti na fázovém poli. Ansiotropie je vyjádřena s pomocí Finslerovy geometrie s dobrými výsledky. Algoritmus využívá MPI a OpenMP k paralizaci napříč vlákny a uzly a vykazuje dobrou škálovatelnost při dvourozměrném dělení třírozměrného prostoru. V první části je vysvětlen model paralení implementace. Druhá část článku je věnována vlivu numerické anisotropie, která je závislá na síti, a dále simulaci komplexního polykrystalického tuhnutí na jemné síti. Všechny simulace jsou prováděny s užitím metody konečných objemů na pravoúhlé pravidelné síti.

Klíčová slova: rovnice fázového pole, metoda konečných objemů, anisotropický růst krystalů

*Tato práce byla podpořena grantem grantové agentury ČVUT v Prague, grant No. SGS17/194/OHK4/3T/14.

Plná verze: P. Strachota, A. Wodecki. *High Resolution 3D Phase Field Simulations of Single Crystal and Polycrystalline Solidification.* Acta Physica Polonica A **134** (2018), 653–657, doi:10.12693/APhysPolA.134.653.

Literatura

[1] M. Beneš. *Diffuse interface treatment of the anisotropic mean-curvature flow.* Appl. Math- Czech. 48, 2003, no. 6, pp. 437–453

[2] R. Eymard. *Finite volume methods, Handbook of Numerical Analysis, vol. 7.* Elsevier, 2000, pp. 715–1022.

[3] P. Strachota. *Analysis and Application of Numerical Methods for Solving Non linear Reaction-Diffusion Equations.* Czech Technical University in Prague Faculty of Nuclear Sciences and Physical Engineering, Dissertation, 2012