DOKTORANDSKÉ DNY 2022

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

11. a 25. listopadu 2022

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2022
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 140, Vydání 1.

Seznam příspěvků

Předmluva

Doktorandské dny jsou již tradičním setkáním studentů doktorského studia na Fakultě jaderné a fyzikálně inženýrské ČVUT v Praze. Doktorandi studijního programu 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 sedmnáctým vydáním workshopu, koná se ve dnech 11. a 25. listopadu 2022. Jsme rádi, že po předchozích letech silně poznamenaných covidem se můžeme opět sejít na workshopu v tradiční prezenční formě.

Tento sborník přináší jak plné texty studentských příspěvků, tak i abstrakty s odkazy na články publikované ve sbornících významných konferencí či publikované nebo alespoň zaslané k publikaci v odborných časopisech.

Za materiální podporu děkujeme katedře matematiky FJFI a grantu Studentské vědecké konference SVK 32/22/F4.

Editoři

Manifold Learning Projection Quality Quantitative Evaluation[*]

Vladislav Belov

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

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

Abstract. A large number of dimensions may cause various problems in real-world applications. Some dimensions might be redundant and can worsen the quality of the workflow output. In the vast majority of exercises with datasets, data are distributed along a highly nonlinear manifold whose structure is unknown. This paper focuses on analyzing the outputs of nonlinear dimensionality reduction, or Manifold Learning, techniques. We introduce three meaningful measures that provide context behind projections onto lower-dimensional spaces. The measures will enable us to compare techniques with each other and assist in choosing suitable hyperparameters. Moreover, we propose to view projections from the standpoint of simplicial complex distortion. In connection with that, we establish the process of a dimension-agnostic graph-based data tessellation technique that builds a simplicial skeleton of high-dimensional data. Alongside our new tessellation technique, we evaluate the proposed quality measures on the Delaunay-tessellation-based simplicial approximations of manifolds.

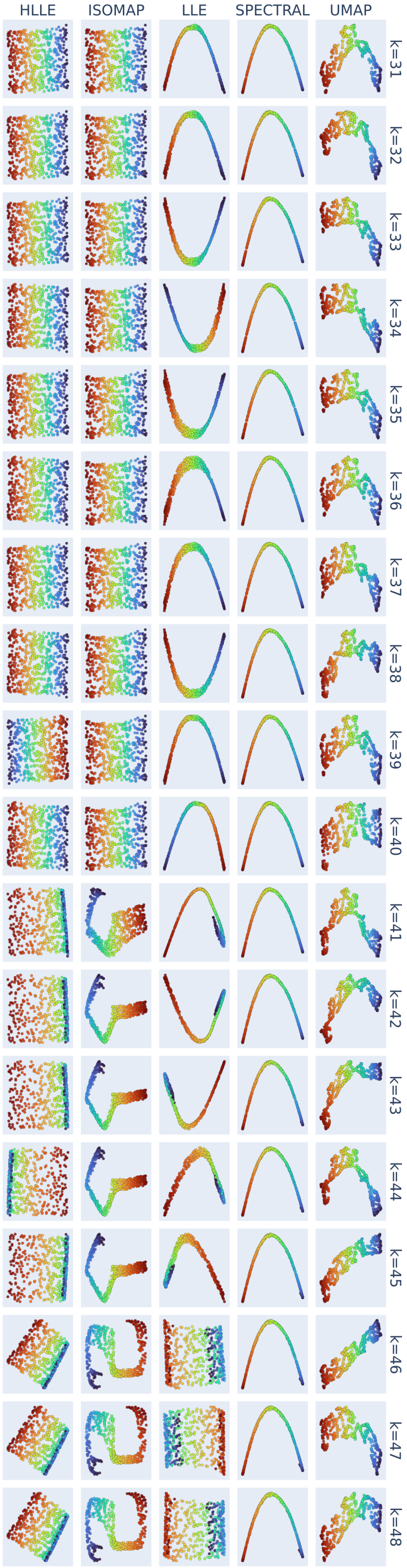
Keywords: dimensionality reduction, machine learning, manifold learning, noise reduction

Abstrakt. Vysoký počet dimenzí může způsobit různé problémy v reálných aplikacích: některé dimenze mohou být redundantní a navíc zhoršující kvalitu výstupů modelů, kterého jsou součástí. Kromě toho, při práci se skutečnými daty se často setkáváme s případy, když jsou distribuována podél nějaké nelineární variety, jejíž struktura je neznámá. Tento příspěvek je zaměřen na analýzu technik nelineární redukce dimenzionality, tzv. Manifold Learning. Zde představujeme tři metriky, které jsou schopné extrahovat informace o prováděných projekcích vysoce dimenzionálních variet na prostory s nižšími dimenzemi. Nadto ukazujeme, že tyto metriky jsou užitečné při výběru jak optimálnější mapovací techniky, tak i jejích hyperparametrů. V této práci také navrhujeme pohled na redukční projekce jako na proces distorze komplexů propojených simplexů. V návaznosti na tuto myšlenku definujeme a užíváme vlastní techniky vytvoření simplexů, která není přímo závislá na dimenzi dat a je založena na principu sestavení simplexové kostry. Při evaluaci navržených metrik také aplikujeme aproximace variet vytvořených na základě Delaunyho teselací.

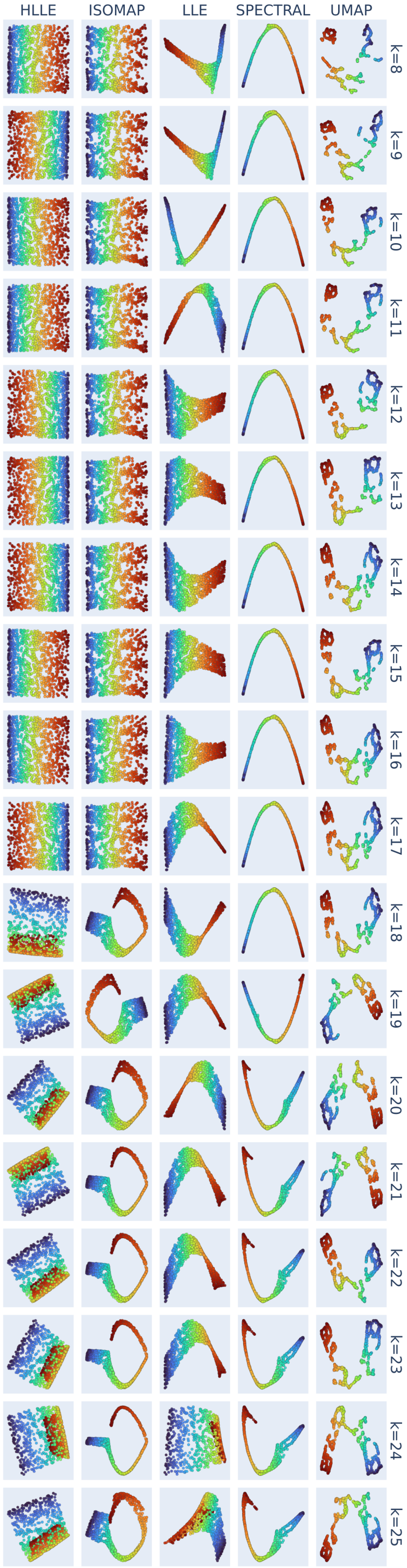
Klíčová slova: Manifold Learning, redukce dimenzionality, redukce šumu, strojové učení

Full paper: The full text of this paper is available in [1]. The work was presented at the CIIS 2021 conference workshop CAIT 2021.

[*] This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS20/183/OHK4/3T/14.



(a) A set of projections of the S-curve, $N = 500$, for different values of the number of nearest neighbors k.



(b) A set of projections of the Swiss Roll, $N = 1000$, for different values of the number of nearest neighbors k.

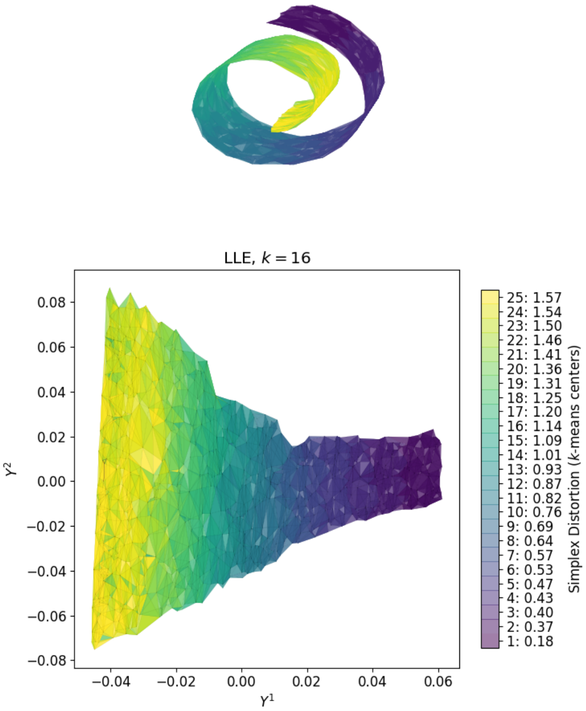


Figure 2: Simplex Distortion of the Swiss Roll dataset, $N = 1000$, DT ($\tau_{\mathrm{DT}} = 3$).

References

[1] V. Belov and R. Marik. Manifold Learning Projection Quality Quantitative Evaluation. In '2021 The 4th International Conference on Computational Intelligence and Intelligent Systems', CIIS 2021, New York, NY, USA, (2021). Association for Computing Machinery.

Tools for Understanding Deep Blind Image Deconvolution*

Antonie Brožová

2nd year of PGS, email: brozoant@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. This paper presents tools that can be used for better understanding of blind image deconvolution algorithms. It is focused on SelfDeblur algorithm, that is based on Deep Image Prior, and aims to uncover some of the reasons of its efficiency using mode-connectivity landscape and power spectral density plots.

Keywords: SelfDeblur, blind image deconvolution, solution landscape, power spectral density

Abstrakt. V tomto příspěvu jsou prezentovány dva nástroje, které lze použít pro lepší pochopení algoritmů řešících slepou dekonvoluci obrazu. Je zaměřen na algoritmus SelfDeblur, který je postavený na myšlence Deep Image Prior, a cílem je lépe porozumět, co stojí za jeho efektivitou, pomocí vykreslení prostoru řešení a grafů výkonové spektrální hustoty.

Klíčová slova: SelfDeblur, slepá dekonvoluce obrazu, prostor řešení, výkonová spektrální hustota

1 Introduction

Images can be degraded in many ways, for example, by blurring, noise or low resolution. In this paper we focus on blurring, which may be caused by a relative motion of a camera and a scene, turbulence in atmosphere or wrong focus. Assuming a spatially invariant blur, a blurred image $\boldsymbol{d} \in \mathbf{R}_{+,0}^{n\times m}$ can be represented as a convolution of a point spread function (PSF) $\boldsymbol{k} \in \mathbf{R}_{+,0}^{s\times s}$ and an underlying sharp image $\boldsymbol{x} \in \mathbf{R}_{+,0}^{n\times m}$

$$\boldsymbol{d} = \boldsymbol{k} \circledast \boldsymbol{x} + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{n} \in \mathbf{R}^{n\times m}$ denotes a noise. The deconvolution is basically an inverse operation to the convolution with the aim to recover the sharp image from the blurred one. The deconvolution is called blind (BID) when not only the sharp image, but also the blur is unknown. The task is then to minimize

$$\|\boldsymbol{d} - \boldsymbol{k} \circledast \boldsymbol{x}\|, \tag{2}$$

with respect to both $\boldsymbol{x}$ and $\boldsymbol{k}$. To preserve the energy, $\boldsymbol{k}$ is required to contain only nonnegative values and sum to 1. This paper is focused on zero-shot blind image deconvolution, which means that the sharp image is estimated without any training on a large dataset.

*This work has been supported by the grant GA20-27939S

1.1 Literature

As the problem is highly ill-posed, some prior information is a vital part of the estimation. Bayesian approach recieved a lot of attention at the begging of the century, starting with Miskin and MacKay [12], Likas and Galatsanos [11] and Molina et al. [13]. Variational Bayes [18], [7] and Maximum Aposteriori (MAP) [9], [14] approaches were mainly discussed and various priors were proposed [20]. Although these traditional methods are quite successful, their efficiency depends on a blur type and inverse operations often leave the sharp images degraded by not very plausible artifacts.

Another interesting approach to the problem of blind image deconvolution is to utilize deep learning [5], [21], [22]. Deep learning models usually require to be trained on large datasets which give them more information than the traditional methods get and, therefore, outperform them. But there are real world scenarios where large datasets are not available, usually because of a screening method, and, for a long time, traditional Bayesian methods were state-of-the-art for these problems. In 2018, Ulyanov et al. proposed Deep Image Prior (DIP) [19] and they state that a structure of a deep neural network is a regularizer of the problem itself and that it may prefer images with certain characteristics. They successfully presented it on image denoising, inpainting, and super-resolution, but not on blind image deconvolution. Ren et al. combined the DIP image network with a fully connected network for the PSF in 2020 and proposed SelfDeblur [15]. This model deblurs image without any training dataset and outperforms the traditional methods that are used for BID.

1.2 SelfDeblur

As described in [15], the model combines two generative neural networks, one for an image, denoted as $\mathcal{G}_x$, and one for a PSF, denoted as $\mathcal{G}_k$. The estimate of the sharp image is generated as $\mathcal{G}_x(\boldsymbol{z}_x)$ and the estimate of the PSF as $\mathcal{G}_k(\boldsymbol{z}_k)$, where $\boldsymbol{z}_x$ and $\boldsymbol{z}_k$ are fixed, randomly sampled arrays from uniform distribution. The deconvolution is then formulated as

$$\begin{gathered}\min_{\mathcal{G}_x,\mathcal{G}_k} \|\boldsymbol{d} - \mathcal{G}_k(\boldsymbol{z}_k) \circledast \mathcal{G}_x(\boldsymbol{z}_x)\|,\\ \text{s.t.} \qquad 0 \leq \mathcal{G}_x(\boldsymbol{z}_x)_i \leq 1, \forall i,\\ \mathcal{G}_k(\boldsymbol{z}_k)_j \leq 0, \forall j, \wedge \sum_j \mathcal{G}_k(\boldsymbol{z}_k)_j = 1. \end{gathered} \tag{3}$$

The requirements of nonnegativity and sum of elements of the PSF can be easily incorporated using softmax and sigmoid output layers. $\mathcal{G}_x$ is as in [19] 5-level U-net [16] with skip connections, batch-normalization, leaky ReLU activations and bilinear upsampling. $\mathcal{G}_k$ is a fully connected neural network with one hidden layer with hardtanh activation. The two networks are optimised jointly in 5000 epochs using Adam optimiser [6] with learning rates 10^{-2} for image and 10^{-4} for blur.

Results in this paper were obtained with a simpler blur model - it is represented only by an array (a bias vector if it was understood as a neural network) and a softmax output layer and it is optimised with a learning rate 10^{-2}, because we mainly focus on

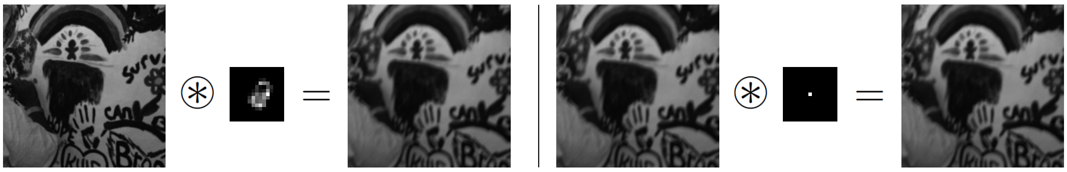


Figure 1: This figure illustrates difference between ground-truth and no-blur solution. On the left side, there is a convolution of ground-truth solution, on the right side, there is a convolution of the no-blur solution. The image and PSF are taken from the Levin dataset.

image network behaviour. Apart from that, random perturbations of $\boldsymbol{z}_x$ are not used and learning rate schedulling is turned off.

2 Studied issues

The problem of BID is highly ill-posed and SelfDeblur suffers from similar problems as traditional methods [8]. One of the main challenges is that there may be inifnitely many solutions minimizing objective (2). Most works focus on how to avoid a so-called no-blur solution, where the PSF is estimated as a dirac delta function and sharp image as the blurred one as shown in Figure 1. To achieve the true sharp image, it is necessary to restrict the solution space. One already mentioned assumption is that the solution is required to preserve energy. Unfortunatelly, this is true for the no-blur solution as well as for the correct one. Therefore, regularization terms, often for both $\boldsymbol{x}$ and $\boldsymbol{k}$ are added to the loss function (2). The authors of SelfDeblur state, that it is enough to minimize the objective without the regularization terms, because the regularization is incorporated naturally by the structure of the image network. In [8] Kotera et al. tested the ability of the network to learn a blurred image and showed that it is actually simpler to train a network generating the blurred image than the sharp one and suggested that the reason of success of SelfDeblur lies more in an optimization method than in the network structure.

Another problem is that deconvolution performed by SelfDeblur results in a different solution every run on the same image. This is caused by the random initialization of weights of the network as well as the input array $\boldsymbol{z}_x$.This means that the starting point has a strong effect on a performance of the deblurring algorithm. Good choice of initialization is an issue shared with traditional Bayesian methods, although they usually start from completely different point than SelfDeblur – the no-blur solution.

Furthermore, good initialization may depend on characteristics of the sharp image. Deconvolution of images in the Levin dataset [10] returns estimates that are way closer to the true sharp ones than estimates of two images from the Kodak dataset [1]. As mentioned by Arican et al. in [2], every sharp image may be represented the best by a different neural network. The U-net is a reasonable choice, but the exact structure returning the best estimate is unknown. There is a lot of decisions to be made: which upsampling method to choose, size of convolutional kernel or number of skip channels. It is generally accepted that the model of the sharp image has to be more accurate than the model of the PSF, because it is assumed to be smaller than the image and, therefore, there is enough data to estimate it well.

Questions that will be discussed in this paper: lies the ability to avoid the no-blur solution only in the representation of the sharp image? Is the ability of an image network to learn a sharp image dependent on characteristics of the sharp image? What role does initialization play in SelfDeblur?

3 Metrics

In this section, two tools for understanding blind image deconvolution will be described.

3.1 Mode connectivity

To explore the solution space of the problem, we decided to utilize a method proposed in [3]. We search for a path between two modes (usually ground truth and no-blur solution) and then project other solutions onto a plane defined by the found path. The path is assumed to be a quadratic Bezier curve and its third control point is found so that an expectation of the loss function w.r.t. uniform distribution on $\langle 0, 1\rangle$ is minimized over the path. Such a landscape is constructed in an image and PSF space together, so that it is possible to compare models that have a different interpretation of parameters. It is rendered as a heatmap of logarithmic values of loss for couples $(\boldsymbol{x}, \boldsymbol{k})$. Other points that are not in this plane are orthogonally projected onto it.

3.2 Power spectral density

Fourier spectra of an image shows its frequency characteristics, which is something that changes when an image gets blurred. To better understand differences between images we decided to use plots of their power spectral densities (PSD). PSD plot is created as a histogram of logarithmic frequency values in the power spectra and these histograms are normalized so that the PSDs can be compared.

4 Experiments

Conducted experiments and their results are presented in this section. Firstly, another model of the sharp image $\boldsymbol{x}$ is described and then compared to results of SelfDeblur with different initial conditions. Eventually, an influence of the U-net structure is discussed. Used images are from the Levin dataset [10] and the Kodak dataset [1], image shown in Figure 3 is used to illustrate behaviour on one sample.

4.1 Architecture of the sharp image model

In classical MAP-based approaches, the image is represented by a matrix and it is assigned a prior distribution that prefers the sharp image to the blurred one. The idea of DIP is that the image prior is a structure of a deep neural network, which is trained to return the sharp image. To test the power of the latent regularization, we decided to compare two image models: sharp image represented only by a matrix and sharp image represented by the U-net. The PSF was represented by an array that sums to one and has nonnegative values. The unknowns were jointly optimized with Adam optimiser [6], which is default for SelfDeblur.

4.2 Initial Conditions

The image network in SelfDeblur is initialized randomly with Kaiming uniform method (default in pytorch). The result of the deblurring is then different every run on the same blurred image. Therefore, we decided to test different initalizations and projected the starts and solutions onto a solution landscape to compare them.

Firstly, we explored whether another way of initialization of the U-net would change the results. The experiment was run on a cutout from an image in Figure 1 and was blurred by 5×5 PSF in the shape of X. We decided to compare Kaiming uniform scheme with Glorot uniform scheme [4] which is often used as well. 35 random initializations from both schemes are projected onto a solution landscape which is rendered in a first row of Figure 2. It can be seen that both initializations tend to cluster closer to the no-blur solution, but Kaiming uniform initializations are way more spread.

Next, an output of randomly initialized U-net was used as an initialization for the matrix image model from section 4.1 to see the influence of the initialization. Optimization was run with both models and the path from initialization to solution was projected onto the landscape (rows 2 and 3 in Figure 2). The paths have quite similar shape and even though the initialization is closer to the no-blur solution in orthogonal projection onto the landscape, it does not get trapped in this minima, and in both cases it converges closer to the ground-truth solution. The simple matrix model does not get as close to the ground-truth solution as SelfDeblur, but converges in the correct direction. PSNR values of images reconstructed in these four runs are in Table 1. It can be seen that the images that are closer to the ground-truth solution in the landscape have higher PSNR values. The difference between the two methods can be seen in Figure 3. Images reconstructed by the simpler matrix model contain ringing artifacts while SelfDeblur is able to avoid this problem. This may be caused by the latent regularization of SelfDeblur, mainly by upsampling layers as will be explained in the next section. It should be noted that the path followed by the simpler matrix model from Kaiming initialization does not correspond with the plotted landscape, so the solution space may be much more complex than this method can show.

On the other hand, when both methods are initialized by a point from the landscape close to the no-blur solution, even SelfDeblur, which was pretrained to start at the point, fails to converge to the correct solution as can be seen in Figure 4. Moreover, when pretrained to start the optimization in a point on a grid on the landscape (only points that contain nonnegative values and psf values sum to one are considered), there are more cases of arriving to the no-blur solution than to the ground-truth solution. It seems that the valley where the ground-truth solution is located is harder to reach. This observation suggests that the random initialization of SelfDeblur is partly responsible for its success when compared to more traditional methods that start from the no-blur solution. Furthermore, starting from a point so different from a real image may act as a warm-up for SelfDeblur and give it enough time to learn more complex connections between the pixels in the sharp image.

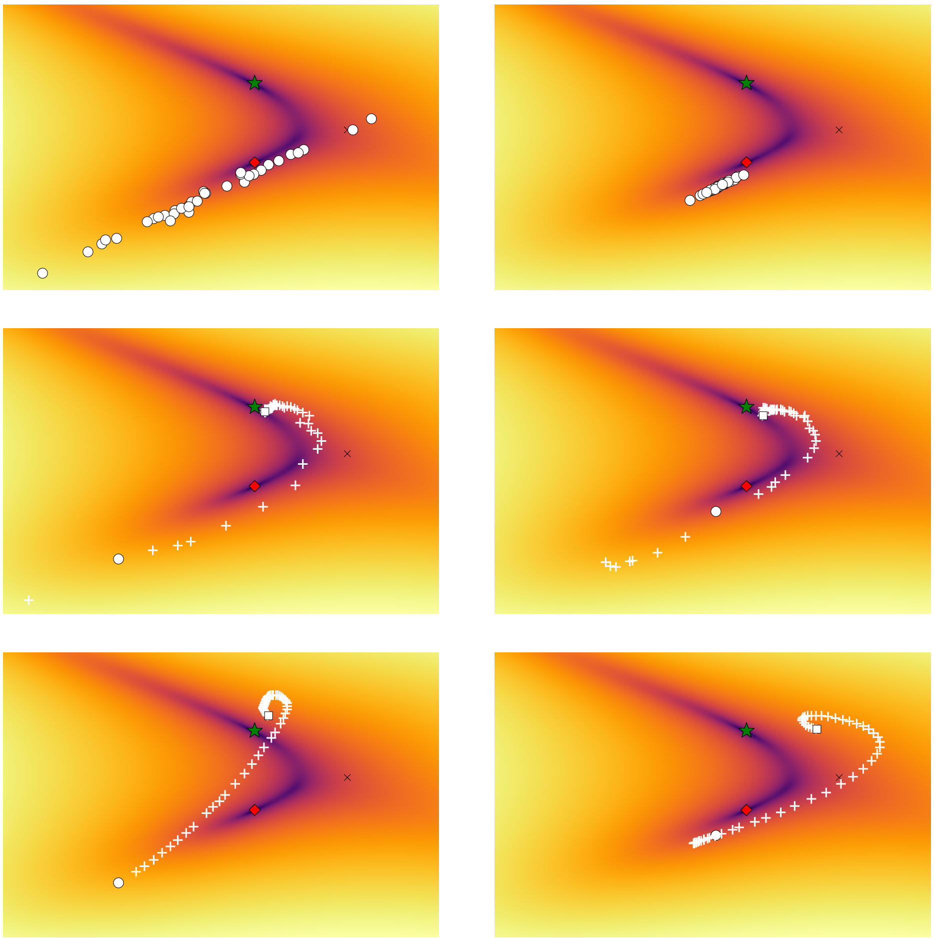


Figure 2: Solution landscape for a cutout from an image in Figure 1. Darker color on the plane shows lower values of loss. Star denotes the ground-truth solution, diamond the no-blur solution and black cross the third control point of the Bezier curve. **First row**: White circles show OG projection of 35 random image network initializations. **Second row**: White crosses show OG projection of path of SelfDeblur from initialization (white circle) to solution (white square). **Third row**: Shows a path of the matrix image representation. **Left** column contains figures of Kaiming uniform initalization, **right** of Glorot uniform initialization.

Table 1: PSNR of reconstructed images from the four paths shown in Figure 2.

	Kaiming	Glorot
SelfDeblur	25.656 dB	26.176 dB
Matrix	23.409 dB	25.239 dB

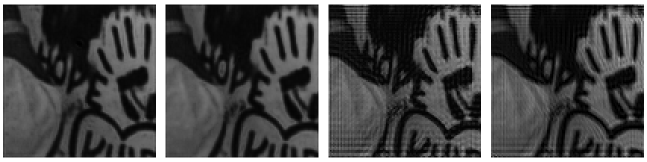


Figure 3: Reconstructed images from the four paths shown in Figure 2. First two are results from SelfDeblur, the other two from the simpler matrix model. First and third were initialized by Kaiming uniform scheme, second and fourth by Glorot uniform scheme.

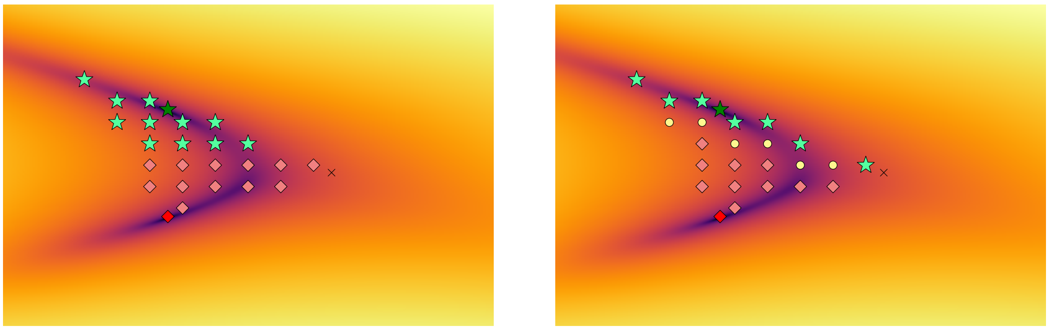


Figure 4: Convergence from points on the solution landscape. **Left** graph shows the results of SelfDeblur, **right** graph the results of the matrix model. Light-colored stars denote points converging to ground-truth solution (darker star), lighter-colored diamonds points converging to the no-blur solution (darker diamond) and circle denotes initializations that did not get close to any of the two modes.

Apart from that, we observed that images with less lower frequency components in Fourier spectrum are harder to deblur (two images from Kodak dataset). Their PSD is plotted in Figure 5. To study the connection to the sharp image, we plotted power spectra of the initializations to see, whether there is any similarity with the power spectra of the sharp image. This hypothesis did not prove to be true. Although power spectra of images initialized according to Glorot and Kaiming scheme are very different and Kaiming's is closer to the sharp images, their efficiencies are similar.

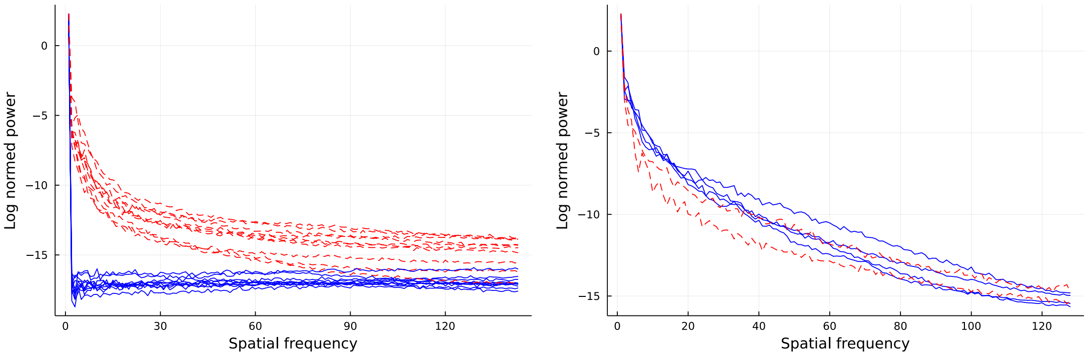


Figure 5: Power spectral densities. **Left** graph shows PSDs of initializations, Glorot uniform by solid line, Kaiming uniform by dashed line. **Right** graph shows PSDs of images from the Levin dataset (solid line) and two images from the Kodak dataset (dashed line).

4.3 Different U-net structures

The U-net image network is composed of convolutional layers, batch-normalization layers, activation functions, skip connections and upsampling and downsampling. Most of these layers have parameters that can be altered so that the network slightly changes and it is possible to compare results produced by these altered networks. The considered parameter changes are a size of convolutional filters (*3* × *3* vs. 5 × 5), variations of a ReLU activation (*LekyReLU(0.2)* vs. ReLU), a number of channels of skip connections (4 vs. *16* vs. 32) and a different upsampling method (*bilinear* vs. nearest neighbour(NN))(parameter values in italics denote original SelfDeblur setting). All these models were were optimised on 222 × 222 cutouts from images from Levin dataset with randomly generated input arrays and initial weights. The deblurring was run 3 times on Levin dataset with each image model variation and PSNR (peak signal-to-noise ratio) of the sharp image estimates are depicted in Figure 6.

All models except for the one with 32 skip channels show slight drop in probability of an image with PSNR around 20dB, which is a value close to PSNR of blurred images. Images with very low PSNR around 16dB are usually images that are not plausible and contain significant artifacts. Most recovered images have PSNR value only slightly lower than 30dB, which is a value that is considered to be a threshold for a good reconstruction. The model with 32 skip skip channels may be able to reconstruct the blurred image easier than the other ones because of a higher ability to copy information throughout the network.

Another difference between models may be visible in power spectra of the reconstructions. Model with NN upsampling returns an image with more higher frequency information than the one with bilinear upsampling as can be seen from an example in Figure 6. Similar observation was also mentioned by Shi et al. in [17], who proposed a method for controlling the amount higher frequency information in a reconstructed image. This may mean that the NN upsampling leads to images that contain more artifacts, but also that it is more suitable for a reconstruction of images with sharper edges and a lot of changes in intensity, while the bilinear upsampling produces smoother images. Considering PSNR, the two modes do not perform significantly different. Other U-net modifications do not produce estimates with any distinguishable differences in power spectra.

5 Conclusion

In this paper, two tools for studying blind image deconvolution algorithms were used to analyze SelfDeblur algorithm. A landscape constructed by mode-connectivity method helped to understand differences in initializations of the sharp image. It was shown that although SelfDeblur starts optimization in a point closer to the no-blur solution than to the ground-truth solution in L_2 norm, it is able to reach the correct sharp solution. Similar behaviour was observed for simpler matrix model of an image, although reconstructed images contained more artifacts. On the other hand, starting optimization from points very close to no-blur solution, it was not able to avoid the undesired minima. One disadvantage of this method is that it is not able to depict a complex landscape, so it can only give an intuition of what a behaviour of an algorithm looks like.

Power spectral density plots showed that there are no obvious connections of a power spectrum of an initialization to a power spectrum of the sharp image, but that images with more lower frequency information may be harder to deblur. Moreover, upsampling method in the U-net influences the result of optimization, namely nearest neighbour

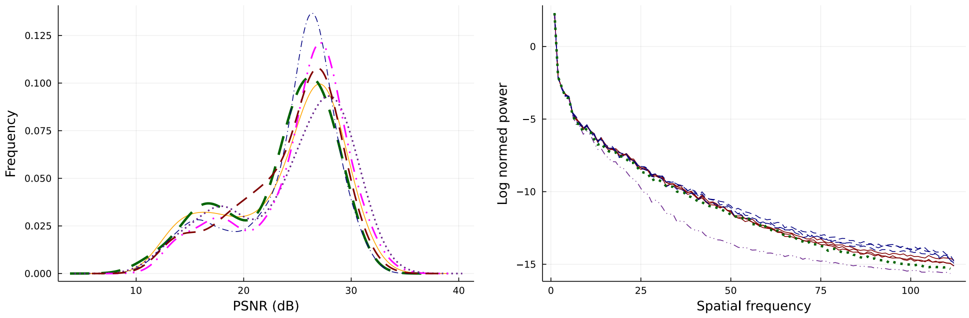


Figure 6: Left: Histograms of PSNR of sharp image estimates from 3 runs on Levin dataset for different U-net sructures. Thicker dashed line was assigned to the original setting of parameters of SelfDeblur, solid line to the NN upsampling, dotted line to the kernel size 5 × 5, dash-and-dotted line to the ReLU activation, dash-and-dot-and-dotted line to 4 skip channels and thinner dashed line to 32 skip channels. **Right**: PSD of reconstructed images with different upsampling modes. Solid line shows bilinear upsampling, dashed line NN upsampling, dotted line PSD of sharp image and dash-and-dot-and-dotted line PSD of the blurred image. The deconvolution was performed on the image from Figure 1.

upsampling returns images with more higher frequency information than bilinear upsampling.

References

[1] Kodak image set. http://r0k.us/graphics/kodak/. Accessed: 2022-04-13.

[2] M. E. Arican, O. Kara, G. Bredell, and E. Konukoglu. Isnas-dip: Image-specific neural architecture search for deep image prior. In '2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', 1950–1958, (2022).

[3] T. Garipov and et al. Loss surfaces, mode connectivity, and fast ensembling of dnns. In 'Advances in Neural Information Processing Systems 31', 8789–8798, Montréal, Canada, (2018). NeurIPS.

[4] X. Glorot and Y. Bengio. *Understanding the difficulty of training deep feedforward neural networks.* Journal of Machine Learning Research **9** (2010).

[5] Y. Huang, E. Chouzenoux, and J.-C. Pesquet. Unrolled variational bayesian algorithm for image blind deconvolution, (2021).

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, (2014).

[7] J. Kotera, V. Šmídl, and F. Šroubek. *Blind deconvolution with model discrepancies.* IEEE Transactions on Image Processing **26** (2017), 2533–2544.

[8] J. Kotera, F. Šroubek, and V. Šmídl. Improving neural blind deconvolution. In '2021 IEEE International Conference on Image Processing (ICIP)', volume 2021, 1954–1958, Anchorage, AK, USA, (2021). IEEE.

[9] A. Levin, W. Yair, D. Fredo, and W. T. Freeman. *Understanding blind deconvolution algorithms.* IEEE Trans Pattern Anal Mach Intell **33** (2011), 2354–2367.

[10] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. *Understanding and evaluating blind deconvolution algorithms.* 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009 (2009), 1964–1971.

[11] A. C. Likas and N. P. Galatsanos. *A variational approach for bayesian blind image deconvolution.* IEEE Transactions on Signal Processing **52** (2004), 2222–2233.

[12] J. Miskin and D. J. C. MacKay. *Ensemble learning for blind image separation and deconvolution.* In 'Advances in Independent Component Analysis', Springer (2000), 123–141.

[13] R. Molina, J. Mateos, and A. K. Katsaggelos. *Blind deconvolution using a variational approach to parameter, image, and blur estimation.* IEEE Transactions on Image Processing **15** (2006), 3715–3727.

[14] D. Perrone and P. Favaro. *A clearer picture of total variation blind deconvolution.* IEEE Trans Pattern Anal Mach Intell **38** (2016), 1041–1055.

[15] D. Ren and et al. Neural blind deconvolution using deep priors. In '2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', 3338–3347, Seattle, WA, USA, (2020). IEEE.

[16] O. Ronneberger, P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation.* In 'Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015', Springer, Cham (2015), 234–241.

[17] Z. Shi, P. Mettes, S. Maji, and C. G. Snoek. *On measuring and controlling the spectral bias of the deep image prior.* International Journal of Computer Vision **130** (4 2022), 885–908.

[18] D. Tzikas, A. Likas, and N. Galatsanos. *Variational bayesian sparse kernel-based blind image deconvolution with student's-t priors.* IEEE Transactions on Image Processing **18** (2009), 753–764.

[19] D. Ulyanov, A. Vedaldi, and V. Lempitski. Deep image prior. In '2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition', 9446–9454, Salt Lake City, UT, USA, (2018). IEEE.

[20] D. Wipf and H. Zhang. *Revisiting bayesian blind deconvolution.* Journal of Machine Learning Research **15** (2014), 3775–3814.

[21] Q. Zhao, H. Wang, Z. Yue, and D. Meng. *A deep variational bayesian framework for blind image deblurring.* Knowledge-Based Systems **249** (2022), 109008.

[22] Z. Zhuang, T. Li, H. Wang, and J. Sun. Blind image deblurring with unknown kernel size and substantial noise, (2022).

Scalable Graph Size Reduction for Efficient GNN Application

Marek Dědič

3rd year of PGS, email: dedicma2@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Martin Holeňa, Department of Adaptive Systems
Institute of Computer Science, CAS
Lukáš Bajer, Cisco Systems, Inc.

Abstract. Graph neural networks (GNNs) present a framework for representation learning on graphs that has been dominant for the past several years. The main strength of GNNs lies in the fact that they can simultaneously learn both from node-related attributes as well as relations between nodes, represented by edges. In tasks leading to large graphs, a GNN often requires significant computational resources to achieve its superior performance. In order to reduce this computational cost, methods allowing for a flexible balance between complexity and performance could be useful. In this work, we propose a simple, scalable, task-aware graph pre-processing procedure that allows us to obtain a reduced graph in such a manner that the GNN achieves a predefined desired performance level on the downstream task in question. In addition, the proposed pre-processing allows for fitting the reduced graph and GNN into given memory/computational resources.

The pre-processing procedure is built on the elementary operation of graph edge contraction. By contracting the edges of the graph one-by-one, a sequence of graphs is obtained, starting with the original one and ending with a graph with no edges and one node per each connected component of the original graph. For each graph in this sequence, a tuple (performance, complexity) can be obtained, where in our work, performance is measured as the accuracy of a classifier on the given downstream task and complexity is measured as the number of nodes in the particular graph. Using these values, the sequence of graphs generated by the pre-processing procedure traces a path in the performance-complexity space. The aim of our work is to study the properties of such a path, with a particular interest in finding a point with the best performance for a given complexity budget, or, conversely, finding the point of lowest complexity for a given required minimal performance.

The edge contraction procedure is driven by an ordering of the edges of the original graph, which in turn defines the aforementioned sequence of graphs. In our work, we define this ordering by measuring the similarity of the predictive posterior distribution of labels of the nodes incident on the edge in question. The choice of a similarity measure is explored experimentally, together with several ways of computing the predictive posterior distribution based on a simplified model specific to the given task.

Additionally, when contracting an edge, a feature aggregation strategy must be defined, as well as a label aggregation strategy. A simple weighted average was used, where the weights are given for the feature aggregation strategy by the number of nodes from the original graph that are represented by a given node. For the label aggregation strategy, similarly, a weighted average of the label distributions of both nodes was used, with the weights representing the number of training nodes represented by a given node. Moreover, when an intermediary graph is to be used for predictions on the original graph, a label refinement strategy is needed in the cases where multiple nodes of the original graph are represented by one node in the intermediary graph. For this strategy, a simple copying of labels was used. This choice of label refinement, however, defines an upper bound on the performance that can be obtained on any given graph in the sequence.

The proposed preprocessing is evaluated and compared with several reference scenarios on conventional GNN benchmark datasets. The performance of the algorithm is compared to the theoretical upper bound defined by the label refinement strategy and the impact of the edge ordering procedure on the performance-complexity characteristics of the algorithm is studied. The main result of this work is that the proposed pre-processing allows for a significant reduction in the number of nodes of a given graph (in some cases, up to 50%) without a major impact on the performance.

Keywords: Graph neural network, Complexity reduction, Hierarchical clustering, Big data

Abstrakt. Grafové neuronové sítě (GNN) představují v posledních letech dominantní nástroj pro reprezentační učení na grafech. Hlavním přínosem GNN je fakt, že se dokáží učit zároveň z příznaků vrcholů a jejich vzájemných vztahů, reprezentovaných hranami. Při řešení úloh vedoucích na velké grafy potřebují GNN často velké množství výpočetních zdrojů aby dosáhly svého vysokého výkonu. Za účelem snížení těchto vysokých výpočetních nároků mohou být užitečné metody dovolující flexibilní kompromis mezi kvalitou předpovědi a výpočetní složitostí. V této práci navrhujeme jednoduchý, škálovatelný, na úloze závisející algoritmus pro předzpracování grafů tak, aby výsledný graf byl zjednodušený takovým způsobem, aby neuronová síť dosáhla požadovaného výkonu na předdefinované cílové úloze. Navrhované předzpracování navíc umožňuje přizpůsobení redukovaného grafu a modelu dostupným výpočetním zdrojům a paměti.

Algoritmus pro předzpracování je postaven na základní operaci kontrakce hrany grafu. Při kontrakci hran jedné po druhé dostáváme posloupnost grafů, počínaje originálním a konče grafem bez hran, kde každá komponenta původního grafu je kontrahována do jednoho vrcholu. Pro každý graf v této posloupnosti lze určit dvojici (výkonnost, složitost). V tomto díle je výkonnost měřena jako přesnost klasifikátoru na dané, předem specifikované úloze a složitost je měřena počtem vrcholů zjednodušeného grafu. Pomocí těchto hodnot lze posloupnost grafů generovanou algoritmem vykreslit jako cestu v prostoru výkonnost-složitost. Cílem této práce je studium vlastností takovéto cesty s mimořádnou pozorností na hledání bodu s nejlepším výkonem pro dané výpočetní možnosti nebo naopak bodu s nejmenší výpočetní složitostí při dosažení daného minimálního výkonu.

Algoritmus zjednodušování grafu je postaven na seřazení hran původního grafu, které skrze jejich kontrakci vyústí ve výše zmíněnou posloupnost grafů. V této práci definujeme takové řazení hran pomocí podobnosti prediktivního posteriorního pravděpodobnostního rozdělení tříd vrcholů přiléhajících na danou hranu. Volba podobnostní míry je zkoumána experimentálně, stejně tak jako několik způsobů výpočtu prediktivního posteriorního pravděpodobnostního rozdělení pomocí jednoduchých modelů specifických pro danou úlohu.

Při kontrakci hran musí navíc být specifikovány strategie pro agregaci příznaků a tříd vrcholů grafu. V této práci bylo použit vážený průměr, kde pro agregaci příznaků byly váhy určeny jako počet vrcholů původního grafu, které jsou reprezentovány daným vrcholem. Pro agregaci tříd byl obdobně použit vážený průměr pravděpodobnostních rozdělení tříd obou vrcholů, kde váhy byly určeny jako počet vrcholů z trénovací sady, které jsou reprezentovány daným vrcholem. V případě, kdy je graf z posloupnosti použit pro klasifikaci na původním grafu je navíc zapotřebí definovat strategii pro zjemňování tříd v situaci, kde jeden vrchol daného grafu odpovídá několika vrcholům původního grafu, potenciálně s různými třídami. Jako tato strategie bylo použité jednoduché kopírování tříd. Tato volba zjemňování tříd definuje horní mez výkonu, který může být dosažený pro daný graf z posloupnosti.

Navrhované předzpracování grafu je vyhodnoceno a porovnáno s několika referenčními scénáři na datasetech běžně využívaných pro vyhodnocování grafových neuronových sítí. Výkon algoritmu je srovnán s teoretickou horní mezí určenou strategií zjemňování tříd a je studován vliv řazení hran grafu na charakteristiku výkonu a složitosti. Hlavním výsledkem této práce je, že navrhované předzpracování grafu umožňuje výrazné snížení složitosti (v některých případech až o 50% vrcholů) bez významnějšího dopadu na výkon.

Klíčová slova: Grafová neuronová síť, Redukce složitosti, Hierarchické shlukování, Big data

Full paper: P. Procházka, M. Mareš, and M. Dědič. Scalable Graph Size Reduction for Efficient GNN Application. In 'Proceedings of the 22nd Conference Information Technologies – Applications and Theory (ITAT 2022)', volume 3226 of *CEUR Workshop Proceedings*, 75–84. `http://ceur-ws.org/Vol-3226/paper9.pdf`

Non-newtonian Turbulent Fluid Flow Simulations through Aortic Phantom Using Cumulant Lattice Boltzmann Method[*]

Pavel Eichler

5th 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 the past decades, computational fluid dynamics (CFD) has begun to be incorporated into clinical examinations as an enhancement of previously used methods. One of the non-invasive methods used so far is the phase-contrast magnetic resonance imaging (PC-MRI) technique. In this case, CFD can be used to eliminate the inaccuracy of PC-MRI, such as low spatial resolution of the acquired data or signal-to-noise ratio. Furthermore, it can be used to provide additional information, where PC-MRI fails to provide reliable data. A typical area where PC-MRI does not provide sufficiently accurate data is the area of turbulent flow, which occurs, for example, in pathologically narrowed areas.

Blood is generally considered a non-Newtonian fluid. The non-Newtonian properties occur mainly in small vessels with diameters approaching the dimensions of the individual blood components. However, in large vessels, blood is typically considered to behave as a Newtonian fluid. It is not yet known whether some non-Newtonian properties of blood play a significant role in flow in pathological areas, e.g. turbulent flow in a narrowed vessel or through a stenotic valve.

One of the key features of the design of the CFD method for the enhancement of PC-MRI data is the computational time, which should not be larger than the measurement time. One option is to use the lattice Boltzmann method (LBM). The advantage of this method is the efficient implementation on graphics cards which can speed up numerical simulations. On the other hand, a non-Newtonian model for the simulated fluid affects the computational requirements even when using LBM.

In this contribution, the effect of Newtonian and non-Newtonian LBM models was investigated using three different fluids and three aortic valves with different severity of stenosis. Numerical simulations were compared with experimental data obtained by PC-MRI. A plastic model (phantom) of the aortic valve with pathological narrowing was used for the experiment. Three fluids were used in the experiment: water, glycerol solution with xanthan gum (GX) and sucrose solution with xanthan gum (SX). The GX and SX fluids represent non-Newtonian fluids with properties similar to human blood.

Based on the severity of the pathological narrowing and the magnitude of flow, the results show that Newtonian models provide comparable results to obtained experimental data, which

[*] The work was supported by the Ministry of Health of the Czech Republic project No. NV19-08-00071,by the Ministry of Education, Youth and Sports of the Czech Republic (MEYS) under the OP RDE grant number CZ.02.1.01/0.0/0.0/16_019/0000765 Research Center for Informatics, and by the project SGS20/184/OHK4/3T/14 of the Student Grant Agency of the Czech Technical University in Prague.

is in favour of overall less expensive Newtonian models.

Keywords: Non-Newtonian fluid, Phase-contrast magnetic resonance imaging, Lattice Boltzmann method, Turbulent fluid flow, Carreau-Yasuda model

Abstrakt. V posledních letech se do klinického vyšetření začala zapojovat výpočetní dynamika tekutin (CFD), jakožto prvek obohacující doposud používané metody. Jednou z doposud používaných neinvazivních metod je měření pomocí magnetické rezonance s aplikací fázového kontrastu (PC-MRI). CFD lze v tomto případě použít k odhalení nedokonalostí této měřící techniky, jako je odstranění nízkého rozlišení získaných dat a odsranění šumu, či k získání dodatečných informací v místech, kde PC-MRI nedokáže poskytnout věrohodná data. Typickou oblastí, kde PC-MRI neposkytuje dostatečně přesná data je oblast turbulentního toku, který se například vyskytuje v patologicky zúžených oblastech.

Krev je obecně považována za nenewtonovskou tekutinu. Tato vlastnost se projevuje převážně v malých cévách o průměrech, které se blíží rozměrům jednotlivých složek krve. Oproti tomu ve velkých cévách se vlastnosti krve spíše podobají vlastnostem newtonovské kapaliny. Není dosud známo, zda při proudění v patologických oblastech, např. při turbulentním proudění v zúžené cévě nebo přes stenotickou chlopeň, nehrají roli některé nenewtonovské vlastnosti krve.

Jedno z klíčových vlastností při návrhu CFD pro doplnění PC-MRI dat je výpočetní čas, který by něměl být příliš větší než je čas samotného měření. Jednou z možností pro CFD je použít metodu mřížkové Boltzmannovy metody (LBM). Výhodou této metody je, že se dá efektivně implementovat na grafických kartách což značně urychlí numerické simulace. Na druhou stranu, použití nenewtonovského podelu pro simulovanou tekutinu ovlivní výpočetní nároky i v případě použití LBM.

V rámci tohoto příspěvku proběhla studie vlivu použití newtonovského a nenewtonovského modelu LBM pro tři různé tekutiny a tři různě vážné patologické zůžení aortální chlopně. Numerické simulace byly srovnány s experimentálními daty získanými pomocí PC-MRI. Pro experiment byl použit plastový model (phantom) aortální chlopně s patologickým zůžením. V experimentu byly celkem použité tři tekutiny: voda, roztok glycerolu s xantanovou gumou (GX) a roztok sacharózy s xantanovou gumou (SX). Tekutiny GX a SX představují nenewtonovské tekutiny a svými vlastnostmi se podobají lidké krvi.

Výsledky ukazují na základě typu vážnosti patologického zůžení a velikosti průtoku, že newtonovské modely poskytují srovnatelné výsledky s experimentálně získanými daty, což je ve prospěch celkově levnějších newtonovských modelů.

Klíčová slova: nenewtonovská tekutina, zobrazení pomocí magnetiské rezonance s aplikací fázového konetrastu, mřížková Boltzmannova metoda, tubrbulentní proudění, Carreaův-Yasudův model.

Full paper: This paper [1] is under review in CAMWA and summarizes results of the research group engaged in project no. NV19-08-00071.

References

[1] Eichler P., Galabov R., Fučík R., Škardová K., Oberhuber T., Pauš P., Tintěra J., Chabiniok R. , Non-Newtonian turbulent flow through aortic phantom: Experimental and computational study using magnetic resonance imaging and lattice Boltzmann method. Under review in CAMWA.

Computation of Feedback Control Laws Based on Switched Tracking of Demonstrations[*]

Jiří Fejlek

5th year of PGS, email: fejlejir@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Stefan Ratschan, Department of Computational Mathematics
Institute of Computer Science, CAS

Abstract. A common approach in robotics is to learn tasks by generalizing from special cases, so-called demonstrations [3]. These are given by a demonstrator [2], for example, in the form of a trajectory optimization method [1]. In this paper, we apply this paradigm to a general control synthesis setting. We present an algorithm that uses such a demonstrator to automatically synthesize a feedback controller for steering ordinary differential equations into a goal set. The resulting control law switches between the demonstrations that it uses as reference trajectories.

This synthesis algorithm constructs the desired controller using a loop that (1) learns a control law, generalizing the current demonstrations to the whole statespace, (2) searches for a counter-example to the desired properties of this control law, and (3) queries the demonstrator for a new demonstration from this counter-example. It iterates this loop until the result is good enough. During this process, it maintains a Lyapunov-like reachability certificate to reduce the simulation time needed in the counter-example search. The algorithm extends construction of control laws based on demonstrations, namely LQR-trees [4, 5], and learning certificates of system behaviour from data/demonstrations [2].

We prove that under some mild assumptions, finitely many cycles of this loop generate a controller that steers the system into the goal set. This is a significantly stronger result than in [4, 5], which only describes the behaviour of the algorithm for the number of iterations tending to infinity. Moreover, we prove that the generated controllers asymptotically reach the performance of the demonstrator.

We also do computational experiments on several examples of dimension up to twelve that demonstrate the practical applicability of the method. We compare our algorithm with controller synthesis fully based on system simulations [4]. In this comparison, our synthesis algorithm runs significantly faster (between 50% and 95%), while producing controllers of similar performance.

Keywords: nonlinear systems, motion planning, learning from demonstrations

Abstrakt. Učení se z názorných příkladů, tzv. demonstrací [3], je běžným přístupem k řešení úloh v robotice. Tyto demonstrace jsou poskytnuty demonstrátorem [2], například v podobě řešiče optimalizace trajektorie [1]. V tomto článku použijeme tento postup pro obecnou úlohu syntézy řízení. Představíme algoritmus, který využívá demonstrace k napočítání zpětnovazebného řízení, které řídí systém popsaný soustavou obyčejný diferenciálních rovnic do dané množiny

[*]This work was supported by the project GA21-09458S of the Czech Science Foundation GA ČR and institutional support RVO:67985807.

cílových stavů. Toto řízení přepíná mezi demonstracemi, které používá jako referenční trajektorie.

Algoritmus napočítává řízení opakováním smyčky, ve které (1) se algoritmus naučí řízení z množiny demonstrací, (2) následně hledá protipříklady s požadovanými vlastnostmi sestavovaného řízení, a (3) závěrem algoritmus doplní množinu demonstrací z nalezených protipříkladů pomocí demonstrátoru. Tato smyčka se opakuje dokud nalezené řízení není dostatečně dobré. V průběhu běhu, algoritmus využívá Ljapunovský certifikát, který umožňuje významně zkrátit simulační čas nutný k vyhodnocení protipříkladů. Algoritmus tak rozšiřuje konstrukci řízení z demonstrací, jmenovitě LQR-stromy [4, 5], a učení se certifikátu z dat/demonstrací [2].

Dokážeme za mírných předpokladů, že algoritmus poskytuje řízení, co dovede systém do množiny cílových stavů, po konečně mnoha iteracích smyčky. Toto je významně silnější výsledek než ten uvedený v [4, 5], který pouze popisuje chování algoritmu pro počet iterací jdoucí do nekonečna. Dále ukážeme, že vygenerované řízení dosahuje asymptoticky chování demonstrátoru.

V článku poskytneme výpočetní experimenty na úlohách až do dimenze dvanáct, které ukazují praktické užití naší metody. Též porovnáme náš algoritmus s algoritmem založeným čistě na simulacích systému [4]. Ukážeme, že náš algoritmus běží znatelně rychleji (o 50% až 95%), přičemž produkuje řízení obdobného výkonu.

Klíčová slova: nelineární systémy, plánování pohybu, učení se z demonstrací

Full paper: Jiří Fejlek and Stefan Ratschan. Computation of Feedback Control Laws Based on Switched Tracking of Demonstrations, arXiv:2011.12639 (`https://arxiv.org/abs/2011.12639`), 2022.

References

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

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

[3] H. Ravichandar, A. Polydoros, C. Sonia, and A. Billard. *Recent advances in robot learning from demonstration.* Annual Review of Control, Robotics, and Autonomous Systems **3** (2020).

[4] 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.

[5] 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.

Efficient Anomaly Detection Through Surrogate Neural Networks

Martin Flusser

7th 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, Avast Software & Institute of Information Theory and Automation, CAS

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

Abstract. Anomaly Detection can be viewed as an open problem despite the growing plethora of known anomaly detection techniques. The applicability of various anomaly detectors can vary depending on the application area and problem settings. Especially in the Big Data industrial setting, an important problem is inference speed, which may render even a highly accurate anomaly detector useless. In this paper, we propose to address this problem by training a surrogate neural network based on an auxiliary training set approximating the source anomaly detector output. We show that existing anomaly detectors can be approximated with high accuracy and with application-enabling inference speed. We compare our approach to a number of state-of-the-art algorithms: one class k-nearest-neighbors (kNN), local outlier factor, isolation forest, auto-encoder, and two types of generative adversarial networks. We perform this comparison in the context of an important problem in cyber-security - the discovery of outlying (and thus suspicious) events in large-scale computer network traffic. Our results show that the proposed approach can successfully replace the most accurate but prohibitively slow kNN. Moreover, we observe that the surrogate neural network may even improve the kNN accuracy. Finally, we discuss various implications that the proposed approach can have while reducing the complexity of applied anomaly detection systems.

Keywords: Anomaly Detector, Neural Network, Model Transfer, Detector Ensemble

Abstrakt. Na detekci anomálií lze pohlížet jako na otevřený problém navzdory rostoucímu množství známých technik detekce. Použitelnost různých detektorů se může lišit v závislosti na oblasti použití a dalších podmínkách. Zejména v průmyslovém prostředí velkých dat je důležitým faktorem rychlost inference, která může způsobit, že i vysoce přesný detektor nebude dobře aplikovatelný. V tomto článku navrhujeme vyřešit tento problém trénováním zástupné neuronové sítě (surrogate neural network) založené na pomocné trénovací sadě aproximující výstup zdrojového detektoru. Ukázali jsme, že stávající detektory anomálií lze aproximovat s vysokou přesností a rychlostí. Tento přístup porovnáváme s řadou nejmodernějších algoritmů jako jsou one class k-nearest-neighbors (kNN), local outlier factor, isolation forest, auto-encoder a dva další typy generativních neuronových sítí. Toto srovnání provádíme v kontextu kybernetické bezpečnosti – odhalování anomálních (a tedy podezřelých) událostí v provozu počítačové sítě. Výsledky ukazují, že navrhovaný přístup může úspěšně nahradit nejpřesnější, ale neúměrně pomalý detektor one-class kNN. Navíc pozorujeme, že náhradní neuronová síť může dokonce zlepšit jeho přesnost.

Závěrem demonstrujeme různé pozitivní důsledky, které navrhovaný přístup přináší zároveň se snižující složitostí aplikace pro systémy detekce anomálií.

Klíčová slova: Detekce anomálií, neuronové sítě, transfer modelu, spojení detektorů

Full paper: M. Flusser and P. Somol. *Efficient anomaly detection through surrogate neural networks.* Neural Computing and Applications (2022), 1–15. URL: `https://rdcu.be/cQKLk`

Mathematical Modeling of the Multicomponent Flow in Porous Media Using Higher-Order Methods[*]

Petr Gális

5th year of PGS, email: galispet@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. In this paper, we present a detailed numerical scheme for a single-phase compressible flow without diffusion of a multi-component mixture in porous media with the higher-order approximation in both space and time. The mathematical model consists of Darcy velocity, transport equations for each component of a mixture, pressure equation and associated relations for physical quantities such as viscosity or equation of state. The discrete problem is obtained using a combination of the discontinuous Galerkin method for the transport equations and the mixed-hybrid finite element method for the Darcy velocity and the pressure equation. In both methods the higher-order approximation is used. The resulting nonlinear problem for concentrations is solved with the fully mass-conservative iterative IMPEC method. Experimental order of convergence analysis (EOC) and some numerical experiments of a 2D flow are carried out.

Keywords: Single-phase flow, Multi-component flow, Discontinuous Galerkin, Mixed-hybrid finite element method, Raviart-Thomas space, Higher-order approximation, IMPEC scheme

Abstrakt. Tato práce se zabývá detailním popisem numerického řešení jednofázového, stlačitelného proudění vícesložkové směsi bez difuze v porézním prostředí pomocí metod vyššího řádu přesnosti v prostoru i čase. Matematický model je popsán Darcyho rychlostí, rovnicí transportu pro každou složku směsi, tlakovou rovnicí a konstitučními vztahy (např. stavová rovnice). K řešení jsme zvolili přístup založený na kombinaci hybridní verze metody smíšených konečných prvků pro řešení tlakového a rychlostního pole a nespojité Galerkinovy metody pro řešení transportních rovnic. Uvedené metody jsou vyššího řádu přesnosti. Výsledný problém je řešen pomocí iterovaného IMPEC schématu. Vyšší řád přesnosti metod je ověřen pomocí experimentální konvergenční analýzy. Použitelnost modelu je ilustrována na několika numerických experimentech.

Klíčová slova: Jednofázové proudění, Proudění vícesložkové směsi, Nespojitá Galerkinova metoda, Hybridní verze metody smíšených konečných prvků, Raviartův-Thomasův prostor, Metoda vyššího řádu přesnosti, Schéma IMPEC

Full paper: Gális P. and Mikyška J. Mathematical modeling of the multicomponent flow in porous media using higher-order methods. Submitted to *Journal of Computational Physics* (2022).

[*]The work has been supported by the projects LTAUSA19021 of the Czech Ministry of Education, Youth, and Sports, and 21-09093S of the Czech Science Foundation.

References

[1] Hoteit H. and Firoozabadi A. Multicomponent Fluid Flow by Discontinuous Galerkin and Mixed Methods in Unfractured and Fractured Media. *Water Resources Research*, 375–391, 2005.

[2] Hoteit H. and Firoozabadi A. Compositional Modeling By the Combined Discontinuous Galerkin and Mixed Methods. *Society of Petroleum Engineers*, 2006.

[3] Chen H. and Fan X. and Sun S. A Fully Mass-Conservative Iterative IMPEC Method for Multicomponent Compressible Flow in Porous Media. *Journal of Computational and Applied Mathematics, Volume 362*, 1–21, 2019.

[4] Moortgat J. and Firoozabadi A. Mixed-hybrid and Vertex-Discontinuous-Galerkin Finite Element Modeling of Multiphase Compositional Flow on 3D Unstructured Grids. *Journal Computational Physics, Volume 315*, 476–500, 2016.

[5] Moortgat J. and Sun S. and Firoozabadi A. Compositional Modeling of Three-Phase Flow With Gravity Using Higher-Order Finite Element Methods. *Water Resources Research, Volume 47*, 2011.

[6] Polívka O. and Mikyška J. Combined Mixed-Hybrid Finite Element–Finite Volume Scheme for Computation of Multicomponent Compressible Flow in Porous Media. *Numerical Mathematics and Advanced Applications*, 559–567, 2011.

[7] Polívka O. and Mikyška J. Numerical Simulation of Multicomponent Compressible Flow in Porous Medium. *Journal of Math for Industry, Volume 3*, 53–60, 2013.

Stochastic Modelling of Fractal Diffusion and Dimension Estimation[*]

František Gašpar

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

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

Abstract. The revision of classical methods for spectral and walk dimension estimates is the main aim of the contribution. Being focused on the unbiased estimation of the walk and spectral dimensions, we aim to construct the estimates with the minimal mean square error. Accompanied simulation experiments are performed on finite substrates, spacial structures serving as a good model of both continuum and fractal sets. We compare the classical approach based on the log-log transform of asymptotic models of returning probabilities and the second moments, and we also develop a weighted approach to improve the statistical properties of dimension estimates. The other discussed aspect is whether to simulate diffusion using the classical graph diffusion model with zero probability of staying in the same vertex or to prefer the physically motivated model of diffusion over edges with the optimal value of jump probability. Finally, we present the results of simulation experiments on two-dimensional finite substrates which approximate the continuum and selected Sierpinski gaskets and carpets. The contribution also summarises general suggestions based on the obtained results from the simulation experiments.

Keywords: diffusion modelling, dimension estimation, fractal substrate, graph representation, spectral dimension, walk dimension.

Abstrakt. Příspěvek si klade se cíl revizi klasických metod pro odhady spektrální dimenze a dimenze náhodné procházky. Důraz je kladen na nestranný odhad dimenze s minimální střední kvadratickou chybou. Doprovodné simulační experimenty jsou provedeny na konečných substrátech, prostorových strukturách sloužící jako dobrý model kontinua i fraktálních množin. Je použit klasický přístup založený na log-log transformaci asymptotických modelů návratových pravděpodobností a druhých momentů. Také je představen vážený přístup ke zlepšení statistických vlastností odhadů dimenze. Dále jsou porovnány simulace difúze pomocí klasického grafového modelu s nulovou pravděpodobností setrvání ve stejném vrcholu a fyzikálně motivovaný model difúze s optimální hodnotou pravděpodobnosti skoku. Jsou představeny výsledky simulačních experimentů na dvourozměrných konečných substrátech, které aproximují kontinuum a vybrané Sierpinského množiny. Příspěvek také shrnuje obecné návrhy pro odhad dimenze na základě získaných výsledků ze simulačních experimentů.

Klíčová slova: difúzní model, odhad dimenze, fraktální substrát, reprezentace grafy, spektrální dimenze, dimenze náhodné procházky.

[*] The authors thank for financial support of grant OPVVV MEYS, Czechia funded project CZ.02.1.01/0.0/0.0/16019/0000765 Research Center for Informatics as well as grant SGS20/190 OHK4/3T/14 of CTU in Prague.

Full paper: F. Gašpar and J. Kukal. *Stochastic modelling of fractal diffusion and dimension estimation.* Physica A: Statistical Mechanics and its Applications 602 (2022), 127624.

An Improved Branch and Bound Algorithm for Phase Stability Testing of Multicomponent Mixtures[*]

Martin Jex

1st year of PGS, email: jexmarti@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. We investigate the phase stability of a multicomponent mixture at constant volume, temperature and moles (VTN stability). Our work is based on the TPD criterion derived in [1] and the branch and bound algorithm from [2]. In this contribution, we improve the algorithm from [2] with more effective bounding strategy. This improvement is achieved using the necessary condition of optimality. In the bounding step of the algorithm, before solving an underestimated convex optimization, we check whether the pressure (given by the Peng-Robinson equation of state) is feasible. If it is not the case, we can exclude the corresponding part of the feasible set from the search. The Peng-Robinson equation of state is not convex and therefore leads to a non convex optimization problem which is computationally expensive. We propose to use a less precise estimate of the global maximum of the pressure. This estimate can be found by comparing the finite number of the values of the tangent plane to a concave overestimate of the Peng-Robinson equation of state. Another benefit of this additional step is to avoid the optimization of the underestimated objective function. The proposed method is tested on several specific examples.

Keywords: phase stability, global optimisation, convex-concave split, branch and bound method, multi component mixtures

Abstrakt. Zkoumáme fázovou stabilitu vícesložkových směsí za konstantního objemu, teploty a molární koncentrace (VTN formulace). Tato práce je založena na kritériu odvozeném v [1] a metodě větví a mezí z [2]. V tomto příspěvku zlepšujeme algoritmus z [2] o lepší zamítaní neperspektvních oblastí přípustné množiny. Tohoto vylepšení je dosaženo s uplatněním nutných podmínek optimality. V kroku mezí, před řešením podhodnoceného konvexního problému, zkontrolujeme, zda je tlak (daný Pengovou-Robinsonovou stavovou rovnicí) přípustný. Pokud tomu tak není, jsme oprávněni tuto část přípustné množiny vyřadit z hledání. Pengova-Robinsonova stavová rovnice není konvexní a tedy je její globální optimalizace výpočetně náročná. Navrhujeme použití méně přesného odhadu globálního maxima tlaku. Tento odhad může být nalezen porovnáním konečného počtu bodů na tečné nadroviny k nadhodnocené konkávní Pengově-Robinsonově stavové rovnici. Další výhoda tohoto kroku je vyhnutí se optimalizace účelové funkce. Metoda je testována na několika příkladech z literatury.

[*] This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic under the OP RDE grant number CZ.02.1.01/0.0/0.0/16 019/0000778 Centre for Advanced Applied Sciences, and by the Czech Science Foundation project no. 21-09093S.

Klíčová slova: dázová stabilita, globální optimalizace, konvexně-konkávní rozklad, metoda větví a mezí, vícesložkové směsi

Full paper: M. Jex, J. Mikyška, An improved branch and bound algorithm for phase stability testing of multicomponent mixtures (2022). Manuscript submitted for publication in Fluid Phase Equilibria.

References

[1] Mikyška, J., Firoozabadi A. *Investigation of Mixture Stability at Given Volume, temperature and number of moles.* Fluid Phase Equlibria, 321, 2012, 1-9.

[2] Smejkal T., Mikyška J. *VTN-phase stability testing using the Branch and Bound strategy and the convex-concave splitting of Helmhotz free energy density.* Fluid Phase Equilibria, 504, 2020, 112323.

[3] Michelsen, M. *The isothermal flash problem. Part I. Stability.* Fluid Phase Equlibria, 9, 1982.

[4] Nagarajan N.R., Cullick A.S., Griewank A. *New strategy for phase equilibrium and critical point calculations by thermodynamic energy analysis. Part I. Stability analysis and flash.* Fluid Phase Equilib, 62, 1991, 191-210.

[5] Nichita D. V., de-Hemptinne J.-C., Gomez S. *Volume-Based Thermodynamics Global Phase Stability Analysis* . Chemical Engineering Communications, 193, 2007, 1194-1216.

[6] Nichita D. V. *Fast and robust phase stability testing at isothermal-isochoric conditions.* Fluid Phase Equilibria, 447, 2017, 107-124.

[7] Nichita D. V. *Volume-based phase stability testing at pressure and temperature specifications.* Fluid Phase Equilibria, 458, 2018, 123-141.

[8] Nichita D. V., Petitfrere M. *Fast and robust phase stability testing at isothermal-isochoric conditions.* Fluid Phase Equilibria, 427, 2016, 147-151.

[9] Smejkal T., Mikyška J. *Unified presentation and comparison of various formulations of the phase stability and phase equilibrium calculation problems.* Fluid Phase Equilib, 476, 2018, 61-88.

[10] Smejkal T., Mikyška J., Kukal J. *Comparison Of Modern Heuristics On Ssolving The Phase Stability Testing Problem.* Discrete Aan Continuous Dynamical Systems Series S, 14, 2021, 1161–1180.

[11] Hartman, P. *On functions representable as a difference of convex functions.* Pacific Journal of Mathematics, 9, 1959, 707-713.

[12] Locatelli M., Schoen F. *Global optimization - Theory, Algorithms, and Applications.* SIAM Philadelphia, 2013.

[13] Boyd S. *Convex Optimization.* Cambridge University Press, 2004.

[14] Jüngel, A., Mikyška, J., and Zamponi, N. *Existence analysis of a single phase flow mixture with van der Waals pressure.* SIAM Journal on Mathematical Analysis, 2018, 50(1): 1367–1395.

Sparsity Techniques in Equation Learning

Zdeněk Junek

2nd year of PGS, email: Zdenek.Junek@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. Deep learning methods are capable to fit complicated structures and provide state-of-the-art results in many domains. However, they are usually dense and over-parametrized and can be pruned significantly with selected sparsification techniques. Moreover, they might provide less accurate results when available training data sets cover only a part of the targeted domain. This can be overcome through a modified model architecture, combined with suitable sparsity technique. Example of such architecture is the so-called equation learning model that is presented in this article. The impact of distinct sparsity techniques employed in this model is experimentally analyzed and compared.

Keywords: Sparsity techniques, equation learning, regularization.

Abstrakt. Metody hlubokého učení dokáží řešit komplikované úlohy v mnoha oblastech. Jsou však zpravidla přeparametrizované s hustými maticemi parametrů, které mohou být významně prořezané za pomoci technik hledání řídkých řešení. Navíc mohou vykazovat horší výsledky, pokud dostupná trénovací data pokrývají pouze část oblasti jejich použití. To může být překonáno pomocí specifické konstrukce modelu kombinované s vhodnou metodou pro hledání řídkých řešení. Jedním z možných řešení je tzv. model učení rovnic, který je prezentován v tomto článku. Hlavní pozornost je věnována vlivu různých použitých metod řídkosti na výstupy tohoto modelu a jejich porovnání.

Klíčová slova: Metody řídkých řešení, učení rovnic, regularizace.

1 Introduction

Deep learning methods have become widely deployed in many domains within last years. They are easily able to fit complicated functions with very large number of parameters.

Nonetheless, classical deep learning models usually suffer from the following drawbacks. First, they are usually dense and over-parametrized and can be pruned significantly without substantial impact on learning accuracy. This over-parametrization brings the need for more computational power and memory, resulting in more costly methods, hence, higher energy consumption required for calculations. Second, they are prone to overfitting which may lead to learning noisy patterns in training data. As a results, these models can show great performance on training data but demonstrate poor results on testing data. Third, many of the models provide "black box" solutions with difficult interpretability of models and their outcomes. All these drawbacks can be addressed by sparsification techniques. Such methods take complex, dense models at the start with the aim to prune such parameters that bring no or negligible additional value to their performance and explanatory power (hence, such models can still utilize complex and flexible model structures as opposed to variable selection techniques which aim to prune input data first and then to adopt simpler models). Sparsity techniques are discusssed in Section 3.

Variety of real-world problems that can be described as regression tasks can be expressed by analytical expression (e.g., mechanical and natural systems), often modeled by complex, non-linear functions. Although such functions may be well approximated by neural networks, it may be beneficial to incorporate a selected set of functions into the model architecture, based on prior knowledge. Such architecture is the key focus of this article and is discussed in Section 2.

Throughout this article, a data set $\mathcal{D}$ with N independent and identically distributed input output pairs is considered: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with regressors $x = \{x_i\}_{i=1}^{N}, x_i \in \mathbb{R}^m$ and targets $y = \{y_i\}_{i=1}^{N}, y_i \in \mathbb{R}^n$. The aim is to find parametrization of a function $h(\cdot|\theta) : \mathbb{R}^m \to \mathbb{R}^n$ in parametrization space Θ such that

$$\mathcal{L}(\theta) = \frac{1}{N}\left(\sum_{i=1}^{N} l\left(h\left(x_i|\theta\right), y_i\right)\right) + \mathcal{R}(\theta)$$

is minimized for $\theta \in \Theta$. Functional forms of the regression function h, the loss function l and the regularization term $\mathcal{R}$ are given as model assumptions. In this article h is either a neural network or its enhancements.

2 Equation learning

Classical neural network-based models usually provide “black box” solutions. However, in some domains like natural sciences, models that provide interpretable results that serve for deeper understanding of a given problem are desired. For instance, it is beneficial to derive a model that can be described by an explicit set of equations, with preferences for simpler equations if possible. Such models often include distinct types of functions including trigonometric functions (e.g., the sinus function in the longitudinal wave equation). Classical neural networks may provide good approximations to such equations; however, they cannot provide a resulting equation unless their architecture is designed for that purpose.

Moreover, in real world application, it may often happen that available data sets cover only a part of the targeted domain. In other words, it is in particular interest not only the question how a model is capable to interpolate data coming from the same distribution, but also how can extrapolate and predict results outside the training range.

To overcome these drawbacks, the so-called “equation learning model” with specifically designed layers was proposed. The pivotal work in this domain is [12]. The notation and architecture in this article is primarily based on this reference. Modifications of this architecture were proposed in [15] with focus on an architecture suitable for equations with division, [2] with a modified architecture inducing a sparse model by construction and [6] with further modifications and applications in a broad range of tasks.

2.1 The model

The model is based on a multilayer feed forward neural network with fully connected layers. Additional calculation units are added to serve for the above described problematics. Let us consider a model with L layers. The base of each layer is a classical fully connected network, with l-th layer as $z^{(l)} = W^{(l)}y^{(l-1)} + b^{(l)}$, where $z^{(l)} \in \mathbb{R}^{k_l}$, $W^{(l)} \in \mathbb{R}^{k_l,d_{l-1}}$ is the weight matrix, $b^{(l)} \in \mathbb{R}^{k_l}$ is the bias vector and $y^{(l-1)} \in \mathbb{R}^{d_{l-1}}$ is the output of the previous $(l-1)$-th layer. For the first layer, the inputs are the regressors $y^{(0)} = x$. The model parameters to be trained are the weight matrices $W^{(l)}$ and the bias vectors $b^{(l)}$ for $l = 1, \ldots L-1$.

Instead of using common activation functions (e.g., the RELU function or the hyperbolic tangent function) to $z^{(l)}$, special transformation functions are introduced in the model for $l = 1, \ldots, L-1$. The first u elements of $z^{(l)}$ are transformed through the so-called "unary units" $f = (f_1, \ldots, f_u)$, univariate functions $f_i : \mathbb{R} \to \mathbb{R}, i = 1, \ldots u$. That is, the first u elements of the l-th layer output $y^{(l)}$ are in the form

$$y_i^{(l)} = f_i\left(z_i^{(l)}\right), \quad i = 1, \ldots, u$$

The number of unary functions u must be taken such that the number of remaining elements $k_l - u$ is even. Selection of the functions may depend on a studied domain and might be based on an expert judgement or a prior knowledge of studied problem. A common selection of the unary units is $f = (sin, cos, \sigma, I)$, where I denotes the identity function and σ the sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$.

The remaining $2v$ elements, $2v = k - u$, are subject to the so-called "binary units" $g = (g_1, \ldots, g_v)$ which are bivariate functions $g_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, i = 1, \ldots v$ resulting in v elements in the layer output $y^{(l)}$

$$y_{u+i}^{(l)} = g_i\left(z_{u+2i-1}^{(l)}, z_{u+2i}^{(l)}\right)$$

. The usual binary functions used in equation learners are the multiplications.

As a result, the layer output $y^{(l)} \in \mathbb{R}^{d_l}$ with $d_l = u + v$ is in the form

$$y^{(l)} = \left(f_1\left(z_1^{(l)}\right), \ldots, f_u\left(z_u^{(l)}\right), g_1\left(z_{u+1}^{(l)}, z_{u+2}^{(l)}\right), \ldots, g_v\left(z_{u+2v-1}^{(l)}, z_{u+2v}^{(l)}\right)\right)$$

.

The last L-th layer is usually taken without the unary and binary functions as $y = \psi\left(W^{(L)}y^{(L-1)} + b^{(L)}\right)$ with ψ being commonly an identity function. Hence, the model has the trainable parameters $\theta = \left(W = \left\{W^{(l)}\right\}_{l=1}^{L}, b = \left\{b^{(l)}\right\}_{l=1}^{L}\right)$.

The model can be trained through usual gradient based techniques (e.g., Adam). The loss function is usually enhanced with a regularization, see Section 3.2.

3 Sparsity techniques in deep learning

There is a broad range of methods and techniques which serve for compression of neural network-based models, like parameter sharing, value quantization or neural architecture search (comprehensive overview of such techniques can be found in [5]). The most common techniques are based on a model pruning that tend to eliminate the weights with low contribution to the final model performance. Substantial part of model pruning methods is based on a regularization or a Bayesian approach (where the former can be in many cases expressed in terms of the latter). Regularization methods are based on an additional regularization term added to the loss function that forces the model parameters to shrink, while the probabilistic Bayesian methods lead to prune the model by introducing sparsity inducing prior distributions of the model parameters into the models (the pivotal works in this area include among others automatic relevance determination [18], dropout-based methods [16], [8], [13], Bayesian group sparsity [9] or prior annealing [17], relation between dropout and shrinkage inducing priors is shown in [14]).

Most common type of a regularization methods is the L_p regularization which adds a regularization term into the loss function $\mathcal{L}(\theta)$ (equation 1) based on the L_p norm of the model parameters

$$\mathcal{R}(\theta) = \lambda\|\theta\|_p$$

Parameter λ governs importance of the regularization term with respect to the error term $\frac{1}{N}\left(\sum_{i=1}^{N} l\left(h\left(x_i|\theta\right), y_i\right)\right)$. The L_p regularization for $p > 0$ tends to shrink the values of the model parameters towards zero unless supported otherwise by the data (hence, this approach is also known as a weight decay). Most common cases in practice are the L_1 regularization, also known as LASSO (from least absolute shrinkage and selection operator), and the L_2regularization, also known as the quadratic regularization or the ridge regression if applied on a regression task. While the L_2 regularization has a tendency to shrink parameters to lower but non-zero values, L_1 may commonly lead to sparser solutions. More detailed discussion on the L_p regularization can be found in [1] and [4], a combination of both was utilized in [19].

3.1 L_0 regularization

Special case of the L_p regularization is for $p = 0$ with the L_0 norm $\|\theta\|_0 = \sum_{i=0}^{M} \chi_{\{\theta_i \neq 0\}}$ that returns the number of non zero elements of θ. Such norm can be a natural way how to force sparsification of the network since the respective regularization term penalizes all non zero weights. Unlike the L_p norm for $p > 0$, the L_0 norm has no tendency to shrink the actual values of the model parameters. It only tends to encourage sparser network parametrizations.

Drawback of the L_0 norm is that it is not differentiable in the parameters, hence, cannot be used in gradient based methods. For this purpose, a way to smooth the norm while retaining its key characteristics was proposed in [10]. This methodology is described in this section. It is assumed the underlying model has M trainable parameters $\theta_i, i = 1, \ldots, M$. Whenever there is a variable indexed by i in this section without a range specification, it is assumed that $i \in \{1, \ldots, M\}$.

First, the so-called binary gates indicating non-zero parameters $z_i \in \{0, 1\}, i = 1, \ldots M$ are introduced into the model. The model parameters can be then rewritten as $\theta_i = \tilde{\theta}_i z_i$. Hence, the i-th parameter is not used in the model if z_i is zero, while $\tilde{\theta}_i$ may remain non-zero. Let us assume that the gates z_i are withdrawn from a $\{0, 1\}$-valued distribution $q(z_i|\pi_i)$ (e.g., the Bernoulli distribution). The objective function can be rewritten as

$$\mathcal{L}\left(\tilde{\theta},\pi\right)=\mathbb{E}_{q(z|\pi)}\left[\frac{1}{N}\left(\sum_{i=1}^{N} l\left(h\left(x_i|\tilde{\theta}\circ z\right),y_i\right)\right)\right]+\lambda\sum_{i=1}^{M}\pi_i$$

where $\circ$ denotes an element-wise multiplication. The aim is to minimize the function in $\tilde{\theta}$ and π.

To smooth the loss function, let us consider a continuous random variable s with a parametric distribution $q\left(s|\phi\right)$. This variable helps to smooth the gates by clipping the variable s in the $[0,1]$ interval

$$z_i=c\left(s\right),\quad c\left(\cdot\right)=\min\left(1,\max\left(0,\cdot\right)\right),\quad s\sim q\left(s|\phi\right)$$

This allows the gate z_i to be exactly zero where $q\left(s|\phi\right)<0$. Moreover, the probability of such event can be easily described by the corresponding cumulative distribution function $Q\left(s\right)$. As a result, the objective function can be rewritten as

$$\mathcal{L}\left(\tilde{\theta},\phi\right)=\mathbb{E}_{q(s|\phi)}\left[\frac{1}{N}\left(\sum_{i=1}^{N} l\left(h\left(x_i|\tilde{\theta}\circ c\left(s\right)\right),y_i\right)\right)\right]+\lambda\sum_{i=1}^{M}\left[1-Q_{s_i}\left(0|\psi_i\right)\right]$$

with the parameters $\tilde{\theta}$ and ϕ to be minimized.

The only condition for the distribution $q\left(s|\phi\right)$ to make the task tractable is that the reparametrization trick (proposed in [7]) can be used for the distribution, i.e., that the distribution $q\left(s|\phi\right)$ can be expressed as a transformation of a simpler non-parametrized distribution $s=T\left(\phi,\varepsilon\right)$ where ε follows a distribution $\varepsilon\sim p\left(\varepsilon\right)$ no longer depending on ϕ. This transformation ensures gradient based methods can be applied since the gradient of the objective function can be expressed as an expectation (since the gradient of the expectation equals the expectation of the gradient if the density $p\left(\varepsilon\right)$ is not a function of ϕ) and can be thus sampled using the Monte Carlo simulation. The objective function becomes

$$\mathcal{L}\left(\tilde{\theta},\phi\right)=\mathbb{E}_{p(\varepsilon)}\left[\frac{1}{N}\left(\sum_{i=1}^{N} l\left(h\left(x_i|\tilde{\theta}\circ c\left(T\left(\phi,\varepsilon\right)\right)\right),y_i\right)\right)\right]+\lambda\sum_{i=1}^{M}\left[1-Q_{s_i}\left(0|\psi_i\right)\right]$$

and is now differentiable w.r.t. $\tilde{\theta}$ and ϕ; the aim remains to find $\tilde{\theta}^{*},\phi^{*}=\underset{\tilde{\theta},\phi}{\operatorname{argmin}}\mathcal{L}\left(\tilde{\theta},\phi\right)$.

As a result, a wide range of distributions can be used. A choice used commonly in practice is the so-called concrete distribution which was introduced specifically as a continuous relaxation of discrete random variables (see [11] for more details). It is a random variable defined on interval $(0,1)$ with two parameters $\phi=(\log\alpha,\beta)$ defined as a transformation of the uniform distribution

$$s=\sigma\left(\frac{\log u-\log\left(1-u\right)+\log\alpha}{\beta}\right),\qquad u\sim\mathcal{U}\left(0,1\right)$$

where σ denotes the sigmoid function. Hence, the reparametrization trick can be used for this distribution by design. The distribution can be expanded to an interval $\left(s_{min},s_{max}\right)$ for $s_{min}<0,\quad s_{max}>1$ as $\bar{s}=s\left(s_{max}-s_{min}\right)+s_{min}$. The gates z_i are then obtained as $z_i=\min\left(1,\max\left(0,\bar{s}_i\right)\right)$.

3.2 Sparsity in equation learning

Since equation learning method aims to find interpretable solutions, sparsity techniques are desired unless model pruning is forced by network architecture by design [2]. Nonetheless, the selection of a suitable sparsity method and comparison of different approaches was not analyzed under scrutiny in the current literature. Most of the methods use the L_p regularization for model pruning with LASSO being the predominant selection. Hence, focus of this article is the comparison of several approaches, assessment if there might be a universally recommended technique for the equation learning tasks and discovery whether the outcomes based on the LASSO regularization may be outperformed (for instance, it is known L_0 regularization may achieve state-of-the-art results in some domains with smaller data sets while performing inconsistently for largescale tasks where comparable or better results may be achieved by simpler methods, [3]).

4 Experiment

The model was tested on the following three equations:

$$\begin{aligned}
EQ1\left(x\right) &= \frac{1}{3}\sin\left(x^{(1)}\right)\\
EQ2\left(x\right) &= \frac{1}{3}\left(\sin\left(\pi x^{(1)}\right)+\sin\left(\pi x^{(2)}+\frac{\pi}{8}\right)+x^{(2)}+x^{(3)}x^{(4)}\right)\\
EQ3\left(x\right) &= \frac{1}{3}\left(\sin\left(\pi x^{(1)}\right)+x^{(2)}\cos\left(2\pi x^{(3)}+\frac{\pi}{4}\right)+x^{(3)}+\left(x^{(4)}\right)^{2}\right)
\end{aligned}$$

with synthetically generated data. The training data was uniformly generated in intervals $x_i \in [-h, h]$, with $x = \left(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}\right)$ in the multivariate cases (EQ2, EQ3), $y = EQ_j(x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2_{dat}\right)$. The following parameters were used for the experiment (unless stated otherwise for specific comparisons): the training set of size $n_{trn} = 10000$ was generated for $h = 1$, and the testing sets (described below) of size $n_{tst} = 3000$ with $\sigma^2_{dat} = 0.01$. The models are based on the equation learning architecture discussed in Section 2, with $L = \{3, 5\}$ layers, using the layers with unary functions $f = (sin, cos, \sigma, I)$ and multiplication as the binary function (i.e., $u = 4$ and $v = 1$). The last layer is always a classical dense layer.

The model parameters were randomly initiated by the glorot distribution. Model was trained by Adam algorithm through 50000 iterations. Sparsity techniques compared are L_2, L_1 and L_0 regularization. They were also compared to the case with no sparsity technique. A default selection of the λ parameter in all regularization algorithms is $\lambda \in \{1\text{e}{-2}, 1\text{e}{-3}, 5\text{e}{-4}, 1\text{e}{-4}, 5\text{e}{-5}, 1\text{e}{-5}, 1\text{e}{-6}\}$, unless other values are needed for further exploration. Parameters used in the L_0 regularization are $s_{min} = -0.1, s_{max} = 1.1$ and $\beta = \frac{2}{3}$. Two testing data sets were created: the interpolation testing set and the extrapolation testing set uniformly generated on the same interval as the training set and on $[-2h, -h] \cup [h, 2h]$, respectively.

First, outcomes for the equation $EQ1$ are demonstrated for the model with $L = 3$ layers in Figure 1. For each method, a Pareto chart is presented with the mean square test error (using the interpolation data set) on the x-axis and a percentage of non-zero parameters on the y-axis (with a tolerance level 1e−3). Curves for L_1 and L_2 regularization

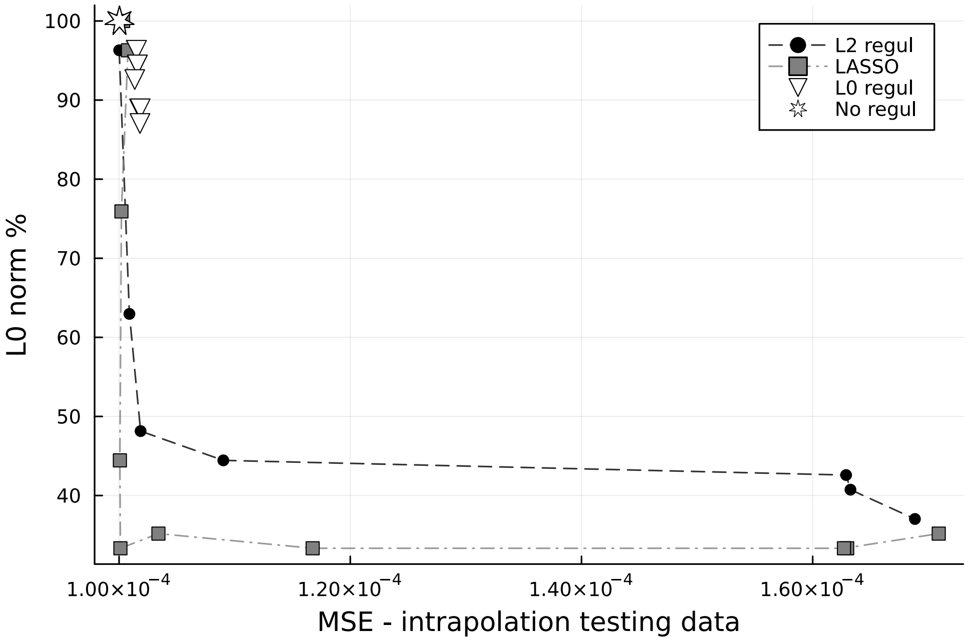


Figure 1: The Pareto chart of the selected techniques on EQ1 (3-layer model), x-axis: the MSE on the interpolation data set, y-axis: the percentage of non-zero model parameters.

follow a "sparsity/accuracy" trade-off (i.e., model accuracy decreases with larger induced sparsity, which happens for increasing λ); however, loss in accuracy for both methods is not substantial while gaining significant sparsity of the trained models. The L_1 regularization prunes larger number of parameters for comparable accuracy and appears most suitable technique for this particular model.

On the contrary, although the L_0 regularization fits the model with large accuracy for a wide range of the values of λ, it demonstrates poor results in terms of sparsity and keeps most of the parameters in the trained model. Even for larger values of λ, there is only a minor betterment in sparsification and it still provides much more dense results than the L_1 and L_2 regularizations while the model accuracy starts to rise dramatically (with the mean square errors outside of the range depicted on the figure).

Conclusions made on the interpolation data set hold also for the extrapolation data set as demonstrated in Figure 2.

Next, the model outcomes were tested for the equations $EQ2$ and $EQ3$ with a more complex model with $L = 5$ layers. The resulting Pareto charts are presented for the interpolation data set in Figure 3. Same behavior as in the previous simpler case is still observed: the L_1 regularization outperforms the L_2 regularization while the L_0 regularization demonstrates poor model pruning. Altogether, the equation learning model appears sensitive to the employed sparsity technique with apparent differences between methods.

5 Conclusion and future work

Importance of model pruning for equation learning was demonstrated and several sparsity techniques compared on selected equations with synthetically generated data. It was

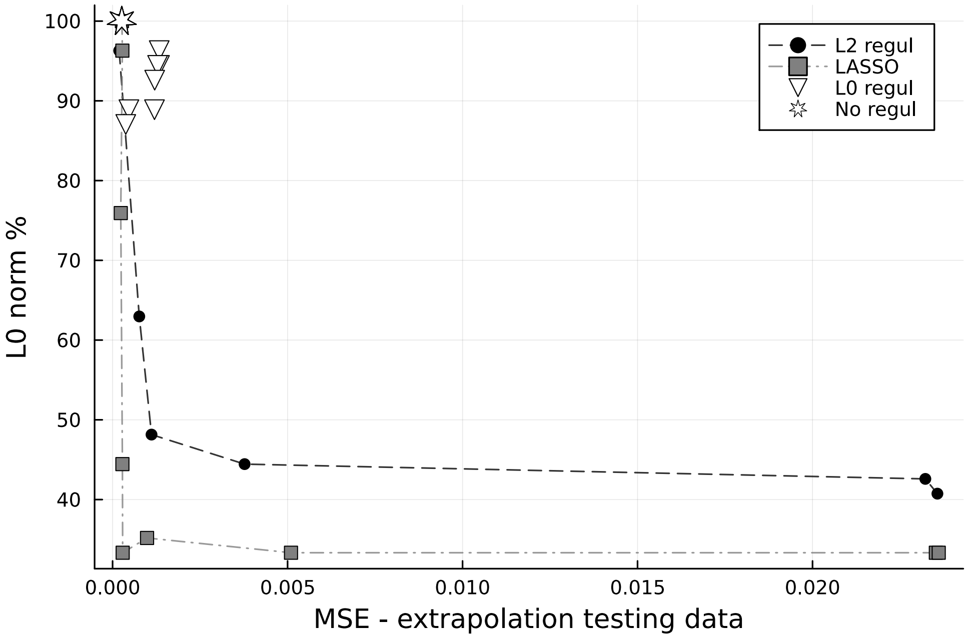


Figure 2: The Pareto chart of the selected techniques on EQ1 (3-layer model), x-axis: the MSE on the extrapolation data set, y-axis: the percentage of non-zero model parameters.

shown that L_1 regularization outperforms other tested methods and provides relatively sparse solution with negligible loss of accuracy compared to case with no regularization.

The goal of future work is assessment of broader range of sparsity techniques on larger number of problematics, including real-world tasks based on real data from several domains with the aim to discover if conclusions made in this article may be generalized to other areas. Also, deeper attention should be paid to the role of sensitivity to random initializations and whether level of sparsity can be still enlarged under specific conditions and utilized techniques.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, (2006).

[2] G. Chen. *Learning Symbolic Expressions via Gumbel-Max Equation Learner Networks*. arXiv:2012.06921 [cs] (May 2021). arXiv: 2012.06921.

[3] T. Gale, E. Elsen, and S. Hooker. *The State of Sparsity in Deep Neural Networks*. arXiv:1902.09574 [cs, stat] (February 2019). arXiv: 1902.09574 version: 1.

[4] S. Han, J. Pool, J. Tran, and W. J. Dally. *Learning both Weights and Connections for Efficient Neural Networks*. arXiv:1506.02626 [cs] (October 2015). arXiv: 1506.02626.

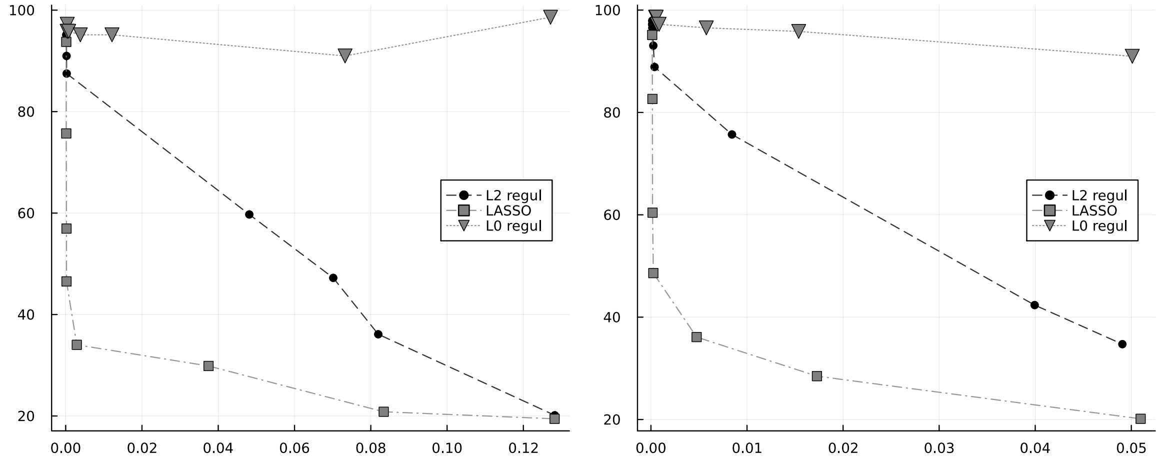


Figure 3: The Pareto chart of the selected techniques on EQ2 (left) and EQ3 (right), 5-layer model, x-axis: the MSE on the interpolation data set, y-axis: the percentage of non-zero model parameters.

[5] T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste. Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks, (January 2021). arXiv:2102.00554 [cs].

[6] S. Kim, P. Y. Lu, S. Mukherjee, M. Gilbert, L. Jing, V. Čeperić, and M. Soljačić. *Integration of Neural Network-Based Symbolic Regression in Deep Learning for Scientific Discovery.* arXiv:1912.04825 [physics, stat] (August 2020). arXiv: 1912.04825.

[7] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes, (May 2014). arXiv:1312.6114 [cs, stat].

[8] D. P. Kingma, T. Salimans, and M. Welling. *Variational dropout and the local reparameterization trick.* Advances in neural information processing systems **28** (2015), 2575–2583.

[9] C. Louizos, K. Ullrich, and M. Welling. *Bayesian compression for deep learning.* arXiv preprint arXiv:1705.08665 (2017).

[10] C. Louizos, M. Welling, and D. P. Kingma. *Learning Sparse Neural Networks through $L_ 0$ Regularization.* arXiv:1712.01312 [cs, stat] (June 2018). arXiv: 1712.01312.

[11] C. J. Maddison, A. Mnih, and Y. W. Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables, (March 2017). arXiv:1611.00712 [cs, stat].

[12] G. Martius and C. H. Lampert. *Extrapolation and learning equations.* arXiv:1610.02995 [cs] (October 2016). arXiv: 1610.02995.

[13] D. Molchanov, A. Ashukha, and D. Vetrov. *Variational Dropout Sparsifies Deep Neural Networks.* arXiv:1701.05369 [cs, stat] (June 2017). arXiv: 1701.05369.

[14] E. Nalisnick, J. M. Hernández-Lobato, and P. Smyth. Dropout as a Structured Shrinkage Prior, (May 2019). arXiv:1810.04045 [cs, stat].

[15] S. Sahoo, C. Lampert, and G. Martius. Learning Equations for Extrapolation and Control. In 'Proceedings of the 35th International Conference on Machine Learning', 4442–4450. PMLR, (July 2018). ISSN: 2640-3498.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. *Dropout: a simple way to prevent neural networks from overfitting*. The journal of machine learning research **15** (2014), 1929–1958.

[17] Y. Sun, W. Xiong, and F. Liang. *Sparse Deep Learning: A New Framework Immune to Local Traps and Miscalibration*. arXiv:2110.00653 [cs, stat] (October 2021). arXiv: 2110.00653.

[18] M. E. Tipping. *Sparse Bayesian Learning and the Relevance Vector Machine*. Journal of Machine Learning Research **1** (2001), 211–244.

[19] H. Yang, W. Wen, and H. Li. *DeepHoyer: Learning Sparser Neural Network with Differentiable Scale-Invariant Sparsity Measures*. arXiv:1908.09979 [cs, stat] (January 2020). arXiv: 1908.09979.

Dual-Cycle: Self-Supervised Dual-View Fluorescence Microscopy Image Reconstruction using CycleGAN*

Tomáš Kerepecký

5th 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. Three-dimensional fluorescence microscopy often suffers from anisotropy, where the resolution along the axial direction is lower than that within the lateral imaging plane. We address this issue by presenting Dual-Cycle (Fig. 1), a new framework for joint deconvolution and fusion of dual-view fluorescence images. Inspired by the recent Neuroclear method, Dual-Cycle is designed as a cycle-consistent generative network trained in a self-supervised fashion by combining a dual-view generator and prior-guided degradation model. We validate Dual-Cycle on both synthetic and real data showing its state-of-the-art performance without any external training data.

Keywords: Light-sheet fluorescence microscopy, Dual-view imaging, deep learning, image deconvolution.

Abstrakt. Trojrozměrná fluorescenční mikroskopie často trpí anizotropií, v jejímž důsledku je rozlišení v axiálním směru nižší než rozlišení ve směru laterárním. Pro řešení tohoto problému jsme navrhli nový framework založený na hlubokém učení (Fig. 1), nazvaný Dual-Cycle, který provádí rekonstrukci dat pomocí spojení dekonvoluce a fúze dvou fluorescenčních 3D obrazů. Náš framework rozšiřuje nedávno publikovanou metodu Neuroclear na data z duálního mikroskopu a přidává apriorní informace o modelu degradace. Dual-Cycle využívá generativní síť s vynucením cyklické konzistence (Cycle-GAN). Trénování je založeno na principu učení bez učitele. Architektura sítě se skládá z generátoru, na jehož vstupu jsou dva 3D obrázky reprezentující tentýž zkoumaný vzorek, ovšem z dvou různých pohledů. Pro cyklickou konzistenci modelujeme degradaci rekonstruovaného obrázku na základě dopředného modelu. Na reálných i synteticky vygenerovaných datech jsme ověřili, že Dual-Cycle dosahuje rekonstrukčních výsledků moderních metod bez využití vnějších trénovacích dat.

Klíčová slova: Light-sheet fluorescenční mikroskopie, duální mikroskop diSPIM, hluboké učení, dekonvoluce obrazu.

Full paper: Tomas Kerepecky, Jiaming Liu, Xue Wen Ng, David W. Piston, and Ulugbek S. Kamilov. arXiv:2209.11729 (https://arxiv.org/abs/2209.11729), 2022.

*This work was supported in part by the Czech Science Foundation grant GA21-03921S, the NSF CAREER award CCF-2043134, the Fulbright commission under the Fulbright-Masaryk award, and by the Beckman Center for Advanced Light-Sheet Microscopy at Washington University in St. Louis.

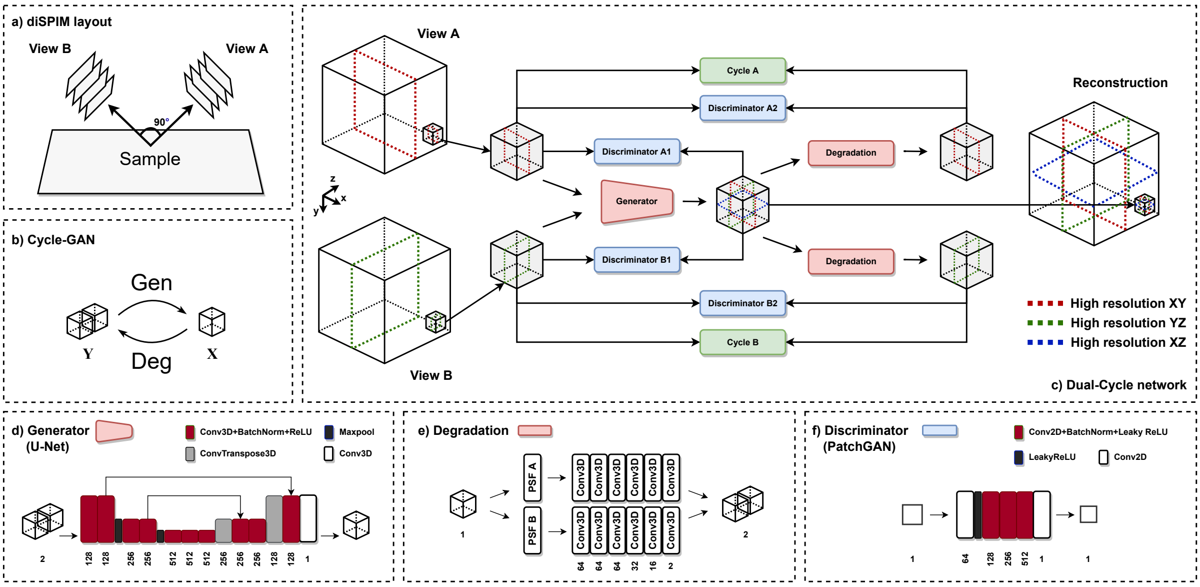


Figure 1: Schematic illustration of the Dual-Cycle framework. a) Scheme of dual-view inverted selective plane illumination microscope (diSPIM). b) CycleGAN approach: for two domains Y and X, CycleGAN learns two mutually-inverse generator mappings *Gen* and *Deg* with the assistance of corresponding discriminators. c) Dual-Cycle network architecture. d) Schematic of the generator based on U-Net. e) Degradation forms two paths each consisting of blurring with known PSF followed by the deep linear generator. f) PatchGAN-based [16] discriminators work on 2D slices of input 3D volumes.

References

[1] E. HK Stelzer, F. Strobl, Bo-Jui Chang, F. Preusser, S. Preibisch, K. McDole, and R. Fiolka, "Light sheet fluorescence microscopy," *Nature Reviews Methods Primers*, vol. 1, no. 1, pp. 1–25, 2021.

[2] R. Liu, Yu Sun, J. Zhu, L. Tian, and U. S Kamilov, "Recovery of continuous 3d refractive index maps from discrete intensity-only measurements using neural fields," *Nature Machine Intelligence*, pp. 1–11, 2022.

[3] William Hadley Richardson, "Bayesian-based iterative method of image restoration," *JoSA*, vol. 62, no. 1, pp. 55–59, 1972.

[4] Leon B Lucy, "An iterative technique for the rectification of observed distributions," *The astronomical journal*, vol. 79, pp. 745, 1974.

[5] Y. Wu, P. Wawrzusin, J. Senseney, R. S Fischer, R. Christensen, A. Santella, A. G York, P. W Winter, C. M Waterman, Z. Bao, et al., "Spatially isotropic four-dimensional imaging with dual-view plane illumination microscopy," *Nature biotechnology*, vol. 31, no. 11, pp. 1032–1038, 2013.

[6] A. Kumar, Y. Wu, R. Christensen, P. Chandris, W. Gandler, E. McCreedy, A. Bokinsky, D. A Colón-Ramos, Zhirong Bao, Matthew McAuliffe, et al., "Dual-view

plane illumination microscopy for rapid and spatially isotropic imaging," *Nature protocols*, vol. 9, no. 11, pp. 2555–2573, 2014.

[7] S. Preibisch, F. Amat, E. Stamataki, M. Sarov, R. H Singer, E. Myers, and P. Tomancak, "Efficient bayesian-based multiview deconvolution," *Nature methods*, vol. 11, no. 6, pp. 645–648, 2014.

[8] M. Temerinac-Ott, O. Ronneberger, P. Ochs, W. Driever, T. Brox, and H. Burkhardt, "Multiview deblurring for 3-d images from light-sheet-based fluorescence microscopy," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1863–1873, 2011.

[9] M. Guo, Y. Li, Y. Su, T. Lambert, D. Dalle Nogare, M. W Moyle, L. H Duncan, R. Ikegami, A. Santella, et al., "Rapid image deconvolution and multiview fusion for optical microscopy," *Nature biotechnology*, vol. 38, no. 11, pp. 1337–1346, 2020.

[10] H. Park, M. Na, B. Kim, S. Park, K. H. Kim, S. Chang, and J. Chul Ye, "Deep learning enables reference-free isotropic super-resolution for volumetric fluorescence microscopy," *Nature Communications*, vol. 13, no. 1, pp. 1–12, 2022.

[11] Y. Wu, X. Han, Y. Su, M. Glidewell, J. S Daniels, J. Liu, T. Sengupta, I. Rey-Suarez, R. Fischer, A. Patel, et al., "Multiview confocal super-resolution microscopy," *Nature*, vol. 600, no. 7888, pp. 279–284, 2021.

[12] J.-Yan Zhu, T. Park, P.Isola, and A. A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[13] "Huygens psf distiller," https://svi.nl/Huygens-PSF-Distiller, Accessed: 2022-08-24.

[14] J. Boutet de Monvel, E. Scarfone, S. Le Calvez, and M. Ulfendahl, "Image-adaptive deconvolution for three-dimensional deep biological imaging," *Biophysical journal*, vol. 85, no. 6, pp. 3991–4001, 2003.

[15] K. Becker, S. Saghafi, M. Pende, I. S.-Litschauer, C. M Hahn, M. Foroughipour, N. Jährling, and H.-U. Dodt, "Deconvolution of light sheet microscopy recordings," *Scientific reports*, vol. 9, no. 1, pp. 1–14, 2019.

[16] P. Isola, J.-Yan Zhu, T. Zhou, and A. A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[17] J. Huisken, J. Swoger, F. Del Bene, J. Wittbrodt, and E. HK Stelzer, "Optical sectioning deep inside live embryos by selective plane illumination microscopy," *Science*, vol. 305, no. 5686, pp. 1007–1009, 2004.

[18] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, et al., "Fiji: an open-source platform for biological-image analysis," *Nature methods*, vol. 9, no. 7, pp. 676–682, 2012.

[19] R. Fernandez and C. Moisy, "Fijiyama: a registration tool for 3d multimodal time-lapse imaging," *Bioinformatics*, vol. 37, no. 10, pp. 1482–1484, 2021.

De Sitter Group in Cosmology*

Jaroslav Kňap

4th year of PGS, email: knapjaro@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. Observations of CMB serve as one of the primary ways to study high-energy gravitational physics. One of the relevant observations is that power spectrum of its fluctuations is nearly scale invariant, as $n_s \approx 0.96$ [1]. This fact suggests that early universe could be well described by a theory possessing scale or conformal symmetry. However, other mundane observations force us to consider that such a symmetry would have to be broken. Natural question is then what would be the remaining symmetry, and common wisdom would be that Poincaré group describes the remaining symmetry.

In this article we make case for consideration of de Sitter group as the symmetry group remaining after spontaneous breakdown of conformal symmetry, and that the appearance of Poincaré group is merely an observational artefact. For that reason we shall discuss general observational and theoretical arguments, supported by a mathematical argument using group contraction.

Keywords: conformal group, de Sitter, group contraction

Abstrakt. Pozorování CMB slouží jako jeden z hlavních nástrojů pro studium gravitační fyziky při vysokých energiích. Jedno z hlavních pozorování je že spektrum jeho fluktuací je takřka škálově invariantní, jelikož $n_s \approx 0.96$ [1]. Tento fakt naznačuje že ranný vesmír je možno vhodně popsat pomocí teorie která má škálovou nebo konformní symmetrii. Další běžná pozorování nás ovšem vedou k závěru že tato symetrie musí být zlomena. Přirozenou otázkou pak je jaká symetrie zůstane po tomto narušení, a přirozenou odpovědí by bylo že Poincareho grupa popisuje tuto symetrii.

V tomto článku předkládáme ke zvážení de Sitterovu grupu jako grupu symetrií která zůstane po spontnáním narušení konformní symetrie, a že zdánlivá přítomnost Poincarého grupy je pouze artefakt pozorování. Za tímto účelem budeme diskutovat obecné pozorovací a teoretické argumenty, s podporou matematického argumentu využívajícího kontrakci grup.

Klíčová slova: de Sitter, konformní grupa, kontrakce group

1 Introduction

One of the current best tests of physics going beyond either Standard Model or General Relativity are astronomical, or more specifically, cosmological observations. Special place among these holds Cosmic Microwave Background, which provides us with a window to an extremely early era in existence of the universe and also to a time of very extreme

*This work has been supported by the grant SGS22/178/OHK4/3T/14

physics. From this observation we can draw a wealth of data, which should provide us with some hints on high-energy physics.

The fact that the power spectrum of CMB is nearly scale invariant [1] hints that when considering inflationary era, it can be considered natural to start from a theory possessing conformal symmetry. We have studied this in previous publications [2], [3], [4], with the resulting observation that the conformal symmetry is spontaneously broken via radiative corrections. We can then ask what is the remnant symmetry that remains after the conformal symmetry is broken. The usual assumptions is that this would be the Poincaré symmetry as that is typically considered as the local space-time symmetry group, however in this article we will make case for considering instead the de Sitter group.

2 Why DeSitter

Before investigating this in more detail, we should engage with a question, why should we consider de Sitter group?

From an observational stand-point we can bring up that detection of 'dark energy' can be considered as an evidence, the effect can be a result of positive cosmology constant [5], which would be present in de Sitter cosmology. Additionally inflationary space-times are also approximately de Sitter [1]. Combined these would suggest that de Sitter group would be more suitable symmetry to consider when describing space-times under going accelerating expansion.

Theoretical arguments [6] also lead us to consider potential introduction of either observer-independent maximum energy or minimal length scales, such theories are called *doubly special relativity.* Major shortcoming of such theories is that they by their very nature require presence of Lorentz violating phenomena at high energies, beyond which we enter a new regime. From this perspective relativistic theory based on de Sitter group (i.e. de Sitter relativity) represents a resolution of this, as it naturally incorporates both invariant velocity and invariant length scale and so Lorentz symmetry remains unbroken and only smoothly transitions to a different high energy regime [7].

Final argument is couched more in an appeal to mathematical 'beauty' and historical precedent.

Until 1905 the prevailing symmetry group considered was the Galilean group, however, the tensions between then new theory of electromagnetism and classical mechanics were resolved that year by A. Einstein in his article '*On the Electrodynamics of Moving Bodies*' [8]. In this article, Einstein overturned and superseded Newtonian mechanics and replaced the Galilean boost with relativistic ones, effectively changing the symmetry group from Galilean to Poincaré.

Very soon after publishing of the article it was intuitively understood that Newtonian mechanics represents low velocity limit of the new theory (or infinite speed of light limit, as the relevant quantity is $\frac{v}{c}$), however despite this understanding there was no proper mathematical method to relate these two theories. On the level of individual formulas it was always possible to perform the expansion and then keep only the lowest relevant order terms, however this was mathematically unsatisfying.

This changed in 1951 when I. Segal published an article '*A class of operator algebras* *which are determined by groups*' [9] where he first proposed idea of a limiting procedure for groups, and its application to symmetry groups of physical theories. These ideas were developed further by E. Inönü and E. P. Wigner who two year later in 1953 published now classical article '*On the contraction of groups and their representations*' [10], where they described specific mathematical method for contracting one group into another. It was immediately clear that this mechanism can be seen as a way to link different theories and to describe theory change.

2.1 Inönü-Wigner group contraction

In this section we quickly summarize the procedure, following similar explanations in [11], [12] and [13].

Since the setting in which Inönü-Wigner contraction operates are continuous symmetries described by Lie groups G_i, the most straightforward way to approach study of this is through investigating transformations of the associated Lie algebras $\mathfrak{g}_i$.

To start of, we have a Lie algebra $\mathfrak{g}$ associated with group G, where J_i is some basis of the vector space of algebra. In this basis we can write commutation relations as

$$[J_i, J_k] = \sum_{k=1}^{n} f_{ijk} J_k \quad i, j = 1, \cdots, n \tag{1}$$

with f_{ijk} being the structure constants of the group. These must satisfy *Jacobi identity*

$$\sum_{c=k}^{n} f_{jlk} f_{ikm} + f_{lik} f_{jkm} + f_{ijk} f_{lkm} = 0. \tag{2}$$

Let us now introduce a *contraction parameter* ϵ and use it to redefine the basis elements $J_i \to J_i^{(\epsilon)}$ such that the following is satisfied:

- the infinite sequence $[J_i]^{\epsilon}$ and its corresponding structure constants $[f_{ijk}]^{(\epsilon)}$ are known
- the limit $\lim_{\epsilon \to 0} [f_{ijk}]^{\epsilon} = [f_{ijk}]^{0}$ exists for all i, j, k and is consistent under under Jacobi identity

If both conditions are satisfied, then the structure constants $[f_{ijk}]^{0}$ generate a new Lie algebra $\mathfrak{g}'$ (and so also Lie group G') called the *contraction* of $\mathfrak{g}$.

It turns out that there are conditions which describe when/how can the contraction be performed for given Lie algebra. Specifically, if $\mathfrak{g}$ has a subalgebra $\mathfrak{h}$ such that

$$\mathfrak{g} = \mathfrak{h} + \mathfrak{p} \tag{3}$$

$$[\mathfrak{h}, \mathfrak{h}] \subset \mathfrak{h}, \quad [\mathfrak{h}, \mathfrak{p}] \subset \mathfrak{p}, \quad [\mathfrak{p}, \mathfrak{p}] \subset \mathfrak{h} + \mathfrak{p} \tag{4}$$

We can then explicitly re-parametrize generators from the subspace $\mathfrak{p}$ as $\mathfrak{p}' = \epsilon \mathfrak{p}$. This does not fundamentally change the algebra, as $\mathfrak{g}_\epsilon$ is isomorphic to $\mathfrak{g}$, since the commutator are

$$[\mathfrak{h},\mathfrak{h}] \subset \mathfrak{h}, \quad [\mathfrak{h},\mathfrak{p}'] \subset \mathfrak{p}', \quad [\mathfrak{p}',\mathfrak{p}'] \subset \epsilon^2 (\mathfrak{h}+\mathfrak{p}) \tag{5}$$

However, if we now perform the limiting procedure $\epsilon \to 0$, the commutators in the singular limit become

$$[\mathfrak{h},\mathfrak{h}] \subset \mathfrak{h}, \quad [\mathfrak{h},\mathfrak{p}'] \subset \mathfrak{p}', \quad [\mathfrak{p}',\mathfrak{p}'] = 0. \tag{6}$$

From this, it is clear that $\mathfrak{g}_0$ is no longer isomorphic to $\mathfrak{g}$, as the subspace $\mathfrak{p}'$ has become an abelian subalgebra after the limit. The algebra $\mathfrak{g}_0$ is clearly the algebra of the contracted group. The structure of the resulting group is that of semi-direct product, i.e. $G' = G_1 \rtimes G_2$ for some groups G_1, G_2.

It is good to notice two things about the procedure:

1. The dimension of the symmetry group is preserved under the contraction. This means that we cannot in this way relate theories that have different number of symmetry generators.

2. The algebra of the contracted group contains an abelian subalgebra, and hence presence of subspace of commuting symmetry generators can be taken as a sign of possible group contraction. On the group level this can be seen from the group structure and presence of semi-direct product.

These observations will be important in the later discussion of symmetries of space-time. We also note that we can assign a geometric interpretation to the contraction parameter, typically as some (pseudo-)radius of the geometry.

Returning to the previous topic, we can note that in the Galilean group $\mathbb{R}^4 \rtimes (\mathbb{R}^3 \rtimes SO(3))$ the boosts form an abelian subalgebra, hinting that the group can be obtained from another via contraction. Indeed, when contracting the Poincaré group $\mathbb{R}^{3,1} \rtimes SO\,(3,1)$ by sending the speed of light to infinity $c \to \infty$ (or equivalently sending the 'slowness' parameter to zero $1/c \to 0$) the Lorentz subgroup transforms as $SO(3,1) \to \mathbb{R}^3 \rtimes SO(3)$. This then provides the firm mathematical link connecting the two theories.

Historically speaking, this relations of the groups, and hence also of Newtonian mechanics and special relativity was only discovered after the theory itself was. However we can also look at it from another angle, that if the mechanism was known prior to formulation of special relativity, we could have speculated whether there is some physical parameter which is sufficiently large (or small), so that effective symmetry group of low velocity mechanics is Galilean, and the full symmetry group is different. Experimentally of course until there was observational evidence for finite invariant speed of light, there was no reason to consider such a parameter to play role in the kinematical group of mechanics.

Taking this second viewpoint we can now notice that Poincaré group $\mathbb{R}^{3,1} \rtimes SO\,(3,1)$ also has an abelian subgroup, in this case the translations. Consequently, we can ask if perhaps there are also not other finite parameters that are currently outside of our observational bounds that we are not considering (e.g. finite invariant length scale), that would lead to Poincaré group being only effective description for large scales. This line of reasoning would lead us to consider de Sitter group (or anti-de Sitter, if we neglect other arguments).

We can succinctly describe this as follows, Poincaré group represented generalization of *high-velocity kinematics* of the Galilean group, and de Sitter group can in turn represents generalization of *high-energy kinematics* of Poincaré group.

3 Symmetries of space-time

With the arguments of the previous section in mind, let us look at the Poincaré group in more detail and derive its relation to de Sitter group. Poincaré group is a 10 dimensional group, which does have the semi-direct product structure, so as stated previously it could be a contraction of another group. It's commutation relations are

$$\frac{1}{\mathrm{i}}\left[M_{\mu\nu}, P_\rho\right] = \eta_{\mu\rho}P_\nu - \eta_{\nu\rho}P_\mu \tag{7}$$

$$\frac{1}{\mathrm{i}}\left[M_{\mu\nu}, M_{\rho\sigma}\right] = \eta_{\mu\rho}M_{\nu\sigma} - \eta_{\mu\sigma}M_{\nu\rho} - \eta_{\nu\rho}M_{\mu\sigma} + \eta_{\nu\sigma}M_{\mu\rho} \tag{8}$$

with

- P_μ - space-time translations
- $M_{\mu\nu}$ - spatial rotations and boosts (spatio-temporal rotations)

The generators $M_{\mu\nu}$ can be explicitly related to the generators of rotations and boosts as $J_i = \frac{1}{2}\epsilon_{ijk}M_{jk}$, $B_i = M_{0i}$. From this we also see that spatial and temporal translations form an abelian subgroup, another hint of the group being potential result of contraction procedure. Specifically, Poincaré group can be obtain as a contraction from either de Sitter group $SO(4,1)$ or anti-de Sitter group $SO(3,2)$, via sending their scalar curvature Λ to zero. For the reasons elaborated on in the previous section, we will be interested in the de Sitter case.

De Sitter group $SO(4,1)$ is once again 10 dimensional group, whose commutation relations can be written succinctly as

$$\frac{1}{\mathrm{i}}\left[M_{AB}, M_{CD}\right] = \eta_{BC}M_{AD} - \eta_{BD}M_{AC} - \eta_{AC}M_{BD} + \eta_{AD}M_{BC}. \tag{9}$$

The contraction to Poincaré group can be constructed as follows, let us define new basis

$$\Pi_\mu = \frac{1}{l}M_{5\mu} \tag{10}$$

with the rest staying the same. Commutation relations can then be rewritten as

$$\frac{1}{\mathrm{i}}\left[\Pi_\mu, \Pi_\nu\right] = \frac{1}{l^2}M_{\mu\nu} \tag{11}$$

$$\frac{1}{\mathrm{i}}\left[M_{\mu\nu}, \Pi_\rho\right] = \eta_{\mu\rho}\Pi_\nu - \eta_{\nu\rho}\Pi_\mu \tag{12}$$

$$\frac{1}{\mathrm{i}}\left[M_{\mu\nu}, M_{\rho\sigma}\right] = \eta_{\mu\rho}M_{\nu\sigma} - \eta_{\mu\sigma}M_{\nu\rho} - \eta_{\nu\rho}M_{\mu\sigma} + \eta_{\nu\sigma}M_{\mu\rho}. \tag{13}$$

Now we can send the pseudo-radius l to infinity, and after that we obtain exactly the commutation relations of the Poincaré group, with $\Pi_\mu \to P_\mu$ becoming the translation generators.

It is clear from the commutation relations that the de Sitter group has neither semi-direct product structure, nor any non-trivial abelian subgroup, so we could in principle end our discussion here.

However, we would like to propose one further step. We start by looking at transformations that can naturally extend de Sitter group, these being operations that act transitively on the de Sitter space [14].

These de Sitter 'translations' can be written as a combination of translations and special conformal transformations, i.e.

$$\partial_\mu - \frac{1}{4l^2}\left(2\eta_{\mu\tau}x^\tau x^\sigma - x^2\delta^\sigma_\mu\right)\partial_\sigma, \tag{14}$$

where l^2 is the de Sitter pseudo-radius, and x_μ some particular coordinates. This object on its own is not de Sitter generator, as its action on space-time leads to conformal rescaling of the metric, transformation which is not from de Sitter group. If we demand the symmetry group of space-time to include transformations under which it is transitive, we must include both the usual translations and the special conformal transformations.

Inclusion of these transformations forces us to also include dilation generator to close the commutation relations of the group. This extensions then moves us to a group $ISO(4,1)$ a 15-dimensional group of isometries of $SO(4,1)$. This group has again the semi-direct product structure, and can be obtained as a contraction of the conformal group $SO(4,2)$, in a similar fashion to the preceding contractions.

There are several reasons to consider conformal group when discussing symmetries of space-time

- conformal group $SO(4,2)$ is the largest symmetry group preserving causality in 4D space-times
- many theories possess scale symmetry (gauge theories, massless theories), which can be promoted to full conformal symmetry under a broad set of conditions (see *Zamolodchikov-Polchinski theorem* in $d = 2$ [15])
- CMB is nearly scale-invariant, suggesting scale invariant (or conformal invariant) theories are suitable for description of very early universe [1]

We consider these arguments to be sufficiently persuasive to explore this direction, so let us take a closer look at conformal group. It's commutation relations can be written as

$$\frac{1}{\mathrm{i}}\left[M_{AB}, M_{CD}\right] = \eta_{BC}M_{AD} - \eta_{BD}M_{AC} - \eta_{AC}M_{BD} + \eta_{AD}M_{BC} \tag{15}$$

in the exact same form as de Sitter ones (which should not be surprising, as both are pseudo-rotation groups). It is more useful to consider an alternative basis that can be related to the generators we are more familiar with

$$L_{\alpha\beta} = M_{\alpha\beta}, \quad D = M_{56} \tag{16}$$

$$P_\alpha = M_{\alpha 5} + M_{\alpha 6}, \quad K_\alpha = M_{\alpha 5} - M_{\alpha 6} \tag{17}$$

$$\alpha, \beta = 0, 1, 2, 3, \quad \alpha = 0 \equiv A = 4. \tag{18}$$

Again, we can see from both the structure of the group, and from the commutation relations that there is no obvious need for further extension. We could again consider transformations acting the space as in de Sitter case, however from a physical perspective we are interested in 3+1 dimensional space-times and any further extension would take us away from that (as $SO(4,2)$ is the conformal group of 3+1 dim space-time).

Additionally, now we are left with a problem, present day universe is well described by considering only 10 space-time symmetries (local), yet conformal group has 15. How can we reduce this number? One answer is spontaneous symmetry breaking, and it turns out that group contraction has a relation to it.

4 Group contraction and symmetry breaking

In this section we describe relation of group contraction to symmetry breaking patters, as developed in [16] and [17].

Great discovery of physics in 20th century was that symmetry in system needs can be realized not just linearly, i.e. in Wigner-Weyl realization, but also non-linearly through so-called Nambu-Goldstone realization. This phenomenon is colloquially known as spontaneous symmetry breaking (SSB) [18], as a particular vacuum state is invariant only under a certain subgroup H of the symmetry group G, i.e. it the symmetry is broken, with the rest of former symmetries now transforming different vacuum states among themselves.

Another well known fact is that spontaneous symmetry breaking leads to appearance of massless Goldstone modes. These fields are either scalar, spinorial or vector, depending on whether the broken symmetry is internal, super or space-time (these can also be composite fields in the case of dynamical symmetry breaking). In relativistic theories the number of Goldstone modes corresponds to the number of broken generators, i.e. to the difference between the dimensions of the symmetry group G and the symmetry group of the vacuum state H,

$$\#\, Goldstone\, modes = \dim G - \dim H \tag{19}$$

It turns out that the 'remnant' symmetry group H combined with the abelian Goldstone modes (which generate field translations) must be a contraction of the original symmetry group G [16]. This then provides us with a mechanism to determine the massless Goldstone modes. Also it provides us with a mechanism to change the dimension of the contracted group, provided we ignore the 'abelianized' field translations.

This relation between symmetry breaking and group contraction gives us a tool to describe the relation of space-time symmetry groups in a unified fashion. We can assume that the conformal group is spontaneously broken (e.g. by introduction of scale, breaking the dilatation generator) and the the left-over non-trivial part is the de Sitter group, with the remaining 5 generators being abelianized and representing field translations.

5 Hierarchy of space-time symmetry groups

Having discussed relation of symmetry breaking and group contraction, we have all the required ingredients ready.

We would propose the following hierarchy of space-time symmetry groups

$$\text{Conformal} \rightarrow \text{de Sitter} \rightarrow \text{Poincare} \rightarrow \text{Galilei}$$

In pre-inflationary and early inflationary universe we propose that the symmetry group of space-time was conformal group $SO(4,2)$. As present day universe is neither conformal nor scale invariant, this part of the symmetry was then spontaneously broken during onset of inflation. The symmetry breaking pattern is then determined by the contraction, as per the previous sections, leading us to de Sitter group $SO(4,1)$

Further we would argue that the true 'local' symmetry of space-time is the de Sitter, not Poincaré. The transition from de Sitter to Poincaré (and from Poincaré to Galilei) are then not true transitions, but merely an approximation artefacts due to limited observational capabilities, of high-velocity physics in once case, and high-energy physics in other.

As the shift to Poincaré happened when we had other theoretical signs and further experimental evidence of finite invariant speed of light, similarly we propose that there is a finite invariant length scale currently out of observational bounds. Existence of this invariant scale would then lead to the symmetry group of space-time being naturally de Sitter, not Poincaré.

Another kind of theoretical argument in support of invariant length (or energy) scale would be that scale on which quantum gravity becomes relevant should play fundamental role. An example could be that a photon of wavelength of Planck length should collapse to black hole (based on classical understanding), however in different reference frames its wavelength could be longer due to red-shift. As a result presence of a black hole would then be observer dependent leading to contradiction within the theory. Eventual quantum theory of gravity should be able to resolve this issue, and treating some length/energy scale as invariant would lead to resolution.

Additionally, we could argue that the structure of Poincaré group combined with knowledge and history of group contraction also points in the direction of further refinement being necessary.

6 Summary and conclusion

In this note we have discussed arguments supporting consideration of de Sitter group as the proper kinematical group of space-time. We have quickly discussed observational evidence supporting consideration of de Sitter group, theoretical argument from doubly special relativity and then moved on to a discussion from the perspective of relating Poincaré and de Sitter groups mathematically.

Primary focus was on application of group contraction to this process, discussing its role in Galilean to Poincaré transition, and by analogy arguing that a similar reasoning can be applied to switch from Poincaré to de Sitter. We have further noted that as de Sitter space time is transitive only under combination of translations and proper conformal transitions, this naturally leads us to consider presence of de Sitter group being the result of spontaneous breakdown of conformal symmetry. As group contraction procedure can be related to symmetry breaking pattern, we can use it to describe that transition also.

We used this to propose that true symmetry group of space-time in early universe was conformal group, which was spontaneously broken down to de Sitter group. Current understanding of Poincaré group as the symmetry group is then an observational artefact resulting from limited experimental and observational capabilities, similar to the way Galilean group was considered as a symmetry of space-time until the developments of tools and other theories caused it to be superseded by Poincaré group.

This allows us to make contact with previous works [2], [3], [4], where we propose conformal theory of gravity for description of inflationary physics. Conformal symmetry present in this theory is spontaneously broken via radiative corrections, with de Sitter group being a good candidate for the resultant symmetry group, for the reasons discussed in this note.

References

[1] Planck Collaboration 2020: *Planck 2018 results. I. Overview and the cosmological legacy of Planck* (A&A 641 A1)

[2] J. Knap: *Quantum Weyl gravity and its cosmological implications, Master thesis*, (CTU Prague, 2019)

[3] P. Jizba, L. Rachwal, J. Knap: *Infrared behavior of Weyl gravity: Functional renormalization group approach*, (Phys. Rev. D 101, 044050, 2020)

[4] P. Jizba, L.Rachwal, S. G. Giaccari, J. Knap: *Dark Side of Weyl Gravity*, (Universe, vol. 6, issue 8, p. 123, 2020)

[5] A. Ashtekar: *Implications of a positive cosmological constant for general relativit*, (Rept.Prog.Phys., 80 (10), pp.102901. 10.1088/1361-6633/aa7bb1. hal-01645576, 2017)

[6] G. Amelino-Camelia: *Doubly Special Relativity*, (Nature 418, 34-35, 2002)

[7] R. Aldrovandi, J. P. Beltran Almeida, J. G. Pereira: *de Sitter special relativity*, (Class. Quant. Grav. 24, 1385-1404, 2007)

[8] A. Einstein: *On the Electrodynamics of Moving Bodies*, (Annalen der Physik 17, 891-921, 1905)

[9] I. E. Segal: *A class of operator algebras which are determined by groups*, (Duke Math. J. 18, 221, 1951)

[10] E. Inönü, E. P. Wigner: *On the Contraction of Groups and Their Representations* (Proc. Natl. Acad. Sci. 39, (6), 510-24 , 1953)

[11] E. Inönü: *Contractions of Lie Groups and their representations*, (Gordan and Breach, Group theoretical concepts in elementary particle physics 391-402, 1964)

[12] R. Gilmore: *Lie Groups, Physics, and Geometry*, (Cambridge University Press, London, 2008)

[13] R. Hermann: *Lie Groups for Physicists*, (W. A. Benjamin, Amsterdam, 1966)

[14] J. G. Pereira, A. Sampson, L. L. Savi: *de Sitter transitivity, conformal transformations and conservation laws*, (Int. J. Mod. Phys. D 23, (4), 2013)

[15] J. Polchinski: *Scale and conformal invariance in quantum field theory*, (Nucl. Phys. B 303, 226-36, 1988)

[16] C. de Concini, G. Vitiello: *Spontaneous breakdown of symmetry and group contractions*, (Nucl. Phys. B 116, 141-56, 1976)

[17] M. Blasone, P. Jizba, G. Vitiello: *Quantum Field Theory and Its Macroscopic Manifestations*, (World Scientific & ICP, London, 2010)

[18] A. J. Beekman, L. Rademaker, J. van Wezel: *An introduction to spontaneous symmetry breaking*, (SciPost Phys. Lect. Notes 11, 2019)

Detection of Higher Order Dependencies in Complex Systems*

Jakub Kořenek

6th 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. The study of complex networks is currently a rapidly developing discipline with applications across various scientific disciplines such as neuroscience, climate research, computer science, economics, energetics, and game theory. A key principle in this area is to view a given system as a network of interacting subsystems (nodes). One of the central questions is to estimate the pattern of their mutual or causal interactions. In this work, we first summarize (often vaguely used) definitions of a total and direct causal effect between two or multiple variables (subsystems) with a particular focus on systems with higher order dependencies, discuss pitfalls of structural causal models of such systems and present a potential information-theoretical concept for determining direct and unique causality.

Keywords: causality, higher order dependencies, mutual information, interventions, XOR function

Abstrakt. Studium komplexních sítí je v současnosti rychle se rozvíjející disciplínou s potenciálními aplikacemi napříč různými vědními obory jako jsou neurověda, klimatologie, informatika, ekonomie, energetika nebo teorie her. Klíčovým principem v této oblasti je nahlížet na daný systém jako na síť vzájemně se ovlivňujících subsystémů (uzlů). Jednou z hlavních otázek je pak odhadnout síť vzájemných kauzálních interakcí mezi těmito uzly. V této práci nejprve shrneme (často vágně používané) definice totálního a přímého kauzálního efektu mezi dvěma nebo více proměnnými (subsystémy) se zvláštním zaměřením na systémy se závislostmi vyšších řádů. Diskutujeme možná úskalí strukturálních kauzálních modelů takových systémů a představíme potenciální informačně-teoretický koncept pro určení přímé a unikátní kauzality.

Klíčová slova: funkce XOR, intervence, kauzalita, vzájemná informace, závislosti vyšších řádů

1 Introduction

The detection of causality is a crucial point in the description of complex systems across scientific disciplines. In practice, we always work with a finite sample of data from which we try to estimate the original causal structure of the system. For this purpose several families of methods like Granger causality or information theoretical approach were suggested but what is the original causal structure? Suppose that we have the exact model (equations) according to which the system behaves - so called structural equation

[*] This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS20/183/OHK4/3T/14 and by the Czech Science Foundation project GA21-17211S.

model. A structural equation model (SEM) (also called a functional model) is defined as a tuple $\mathcal{S} := \left(\mathbf{S}, \mathbb{P}^{\mathbf{N}}\right)$, where $\mathbf{S} = (S_1, \ldots, S_n)$ is a collection of n equations

$$S_j : X_j = f_j \left(\mathbf{PA}_j, N_j\right), \quad j = 1, \ldots, n, \tag{1}$$

where $\mathbf{PA}_j \subseteq \{X_1, ..., X_n\} \setminus \{X_j\}$ are called parents of X_j and $\mathbb{P}^{\mathbf{N}} = \mathbb{P}^{N_1,\ldots,N_n}$ is the joint distribution of the noise variables, which we require to be jointly independent, i.e., $\mathbb{P}^{\mathbf{N}}$ is a product distribution. The graph of a structural equation model is obtained simply by drawing direct edges from each parent to its direct effects, i.e., from each variable X_k occurring on the right-hand side of equation (1) to X_j. We also say that SEM $\mathcal{S} := \left(\mathbf{S}, \mathbb{P}^{\mathbf{N}}\right)$ is generating distribution of $\mathbf{X} = (X_1, \ldots, X_n)$ $\mathbb{P}^{\mathbf{X}}$. But this graph does not have to be a causal graph - an intuitive counter-example is the function $X_1 = 0.X_2$ where of course variable X_2 does not have any causal effect on X_1. We define causality due to Judea Pearl [2] using so-called interventional distribution. Consider a distribution $\mathbb{P}^{\mathbf{X}}$ that has been generated from an SEM $\mathbf{S} = (S_1, \ldots, S_n)$. We can then replace one (or more) structural equations (without generating cycles in the graph) and obtain a new SEM $\tilde{\mathcal{S}}$. We call the distributions in the new SEM interventional distributions and say that the variables whose structural equation we have replaced have been "intervened on". We denote the new distribution by

$$\mathbb{P}^{\mathbf{X}}_{\tilde{\mathcal{S}}} = \mathbb{P}_{\mathcal{S}}^{\mathbf{X}|do\left(X_j=\tilde{f}\left(\tilde{\mathbf{PA}}_j,\tilde{N}_j\right)\right)}. \tag{2}$$

The set of noise variables in $\tilde{\mathcal{S}}$ now contains both some "new" $\tilde{N}$'s and some "old" N's and is required to be mutually independent. The causal effect is then defined as follows. Given an SEM $\mathcal{S}$, there is a (total) causal effect from X to Y if and only if there is x, such that $\mathbb{P}_{\mathcal{S}}^{Y|do(X=x)} \neq \mathbb{P}_{\mathcal{S}}^{Y}$. Note that we can easily define causal effect generally from set of variables $\mathbf{X}$ to Y by replacing $do\,(X = x)$ by $do\,(\mathbf{X} = \mathbf{x})$. As Pearl declares, in practice, we are unable to determine causality (or the direction of causality) without interventions. In certain situations, however, we can assume that some variables, so-called source variables, can influence the so-called target variable but not vice versa. One of these situations is time-ordered data, where only things in the past can affect things in the present. Suppose source variables $(X_1, \ldots, X_n)$ and target variable Y, we say that X_i has causal effect on Y if

$$I\left(X_i, Y\right) > 0. \tag{3}$$

Also, we can say that there is a direct causal effect from X_i to Y if

$$I\left(X_i, Y \middle| \mathbf{X} \setminus \{X_i\}\right) > 0, \tag{4}$$

however, as we show later, this definition of a direct causal effect is valid only for first-order dependencies and collapses for higher-order ones.

2 Higher order dependencies

There is no single definition of synergy or second-order dependence but we show this concept on the well-known example of XOR function whose values are defined according to the table 1. Simply, we consider two fair coins X_1 and X_2, if the result of their flips is the same then the value of the XOR function is zero otherwise it is one.

$$\begin{aligned} &X_1, X_2 \overset{iid}{\sim} Be(0.5) \\ &Y = XOR\left(X_1, X_2\right). \end{aligned} \tag{5}$$

This system is a model example of so-called synergy, there is no information of Y in X_1 neither in X_2, but we get the information about Y from tuple (X_1, X_2). As both variables X_1 and X_2 appear on right hand side of equation (5), the structure of graph of SEM is $X_1 \to Y \leftarrow X_2$ but no matter how we intervene on X_1 or X_2, distribution of Y remain unchanged because as X_1 and X_2 can reach only value of 0 or 1, all possible intervention are in form of Bernoulli distribution $Be(p)$. Suppose that we intervene on $X_1 \sim Be\left(p\right)$ then

X_1	X_2	$Y = XOR\left(X_1, X_2\right)$	p
0	0	0	1/4
0	1	1	1/4
1	0	1	1/4
1	1	0	1/4

Table 1: Function $XOR(X_1, X_2)$

$$\begin{aligned} p(Y = 0) &= p(X_1 = 0, X_2 = 0) + p(X_1 = 1, X_2 = 1) \\ &= (1 - p) * 1/2 + p * 1/2 = 1/2 \end{aligned} \tag{6}$$

$$\begin{aligned} p(Y = 1) &= p(X_1 = 0, X_2 = 1) + p(X_1 = 1, X_2 = 0) \\ &= (1 - p) * 1/2 + p * 1/2 = 1/2, \end{aligned} \tag{7}$$

hence

$$\mathbb{P}^{Y}_{\mathcal{S}} \sim Be\left(0.5\right) = \mathbb{P}^{Y|do(X_1 = Be(p))}_{\mathcal{S}} = \mathbb{P}^{Y|do(X_2 = Be(p))}_{\mathcal{S}}. \tag{8}$$

Only when intervene on tuple (X_1, X_2) we change distribution on Y, for example

$$\mathbb{P}^{Y}_{\mathcal{S}} \sim Be\left(0.5\right) \neq Be\left(1\right) \sim \mathbb{P}^{Y|do((X_1, X_2) = (0,1))}_{\mathcal{S}}. \tag{9}$$

Thus, the causal structure cannot be captured by a graph, but only by a hypergraph. Using an information theoretical approach to direct causality, we however end up with a curious result. As we already proposed, both mutual information are equal to zero (eq. (11)) but conditional mutual information $I(X_1, Y|X_2)$ and $I(X_2, Y|X_1)$ are equal to one and therefore there is a direct causal effect from both of the variables by definition. For this reason, we redefine information theoretical based direct causal effect in accordance with Pearl's definition.

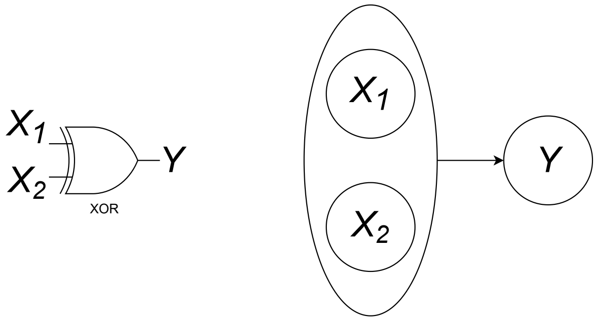


Figure 1: Scheme and causal hypergraph of XOR system

$$I(X_1, Y) = I(X_2, Y) = 4 * \frac{1}{4} \log \frac{1/4}{1/2\ 1/2} = 0\text{bit} \tag{10}$$

$$I((X_1, X_2), Y) = 4 * \frac{1}{4} \log \frac{1/4}{1/4\ 1/2} = 1\text{bit} \tag{11}$$

$$I(X_1, Y|X_2) = I((X_1, X_2), Y) - I(X_2, Y) = 1\text{bit} \tag{12}$$

$$I(X_2, Y|X_1) = I((X_1, X_2), Y) - I(X_1, Y) = 1\text{bit}. \tag{13}$$

2.1 Information-theoretical approach to direct causality

First, we remind Pearl's approach to direct causality using interventions. Given an SEM $\mathcal{S}$, there is a direct causal effect from X_i to Y if and only if there are x_i, $\tilde{x}_i$ such that for every possible values $\mathbf{x} \smallsetminus \{x_i\} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$

$$\mathbb{P}_{\mathcal{S}}^{Y|do(X_i=x_i;\mathbf{X}\smallsetminus\{X_i\}=\mathbf{x}\smallsetminus\{x_i\})} \neq \mathbb{P}_{\mathcal{S}}^{Y|do(X_i=\tilde{x}_i;\mathbf{X}\smallsetminus\{X_i\}=\mathbf{x}\smallsetminus\{x_i\})}. \tag{14}$$

Because of the inconsistency in the information theoretical approach described above, we suggest a new information-theoretical definition of a direct causal effect. Let $\mathbf{X} = (X_1, \ldots X_n)$ be a set of source variables and Y a target variable, there is a direct causal effect from X_i to Y if and only if

$$I(X_i, Y|\mathbf{S}) > 0 \tag{15}$$

for all $\mathbf{S} \subseteq \mathbf{X} \smallsetminus \{X_i\}$. Note that conditioning by empty set is meant standard mutual information

$$I(X_i, Y|\emptyset) = I(X_i, Y). \tag{16}$$

Generally, we can define direct causal effect between set of variables and one target variable as: Let $\mathbf{X} = (X_1, \ldots X_n)$ be a set of source variables and Y a target variable, there is a direct causal effect from $(X_{i_1}, \ldots, X_{i_k})$ to Y if and only if

$$I((X_{i_1}, \ldots, X_{i_k}), Y|\mathbf{S}) > 0 \tag{17}$$

for all $\mathbf{S} \subseteq \mathbf{X} \smallsetminus \{X_{i_1}, \ldots, X_{i_k}\}$. Note that if there is direct causal effect from variable X_i then there is also direct causal effect from any set containing this variable. For determine the "true" unique causal direct effect we can use partial information decomposition [1].

2.2 Example: Indirect $XOR(X_1, X_2)$

We show the consistency of our approach and Pearl's on an example of undirect XOR link and direct linear link given by SEM:

$$\begin{aligned} X_1, X_2 &\overset{iid}{\sim} Be(0.5) \\ X_3 &= XOR(X_1, X_2) + \mathcal{E}_3 \\ Y &= X_3 + \mathcal{E}_Y, \end{aligned} \tag{18}$$

where $\mathcal{E}_3, \mathcal{E}_Y \overset{iid}{\sim} Be(0.5)$.

X_1	X_2	$\mathcal{E}_3$	$\mathcal{E}_Y$	$XOR(X_1, X_2)$	X_3	Y	p
0	0	0	0	0	0	0	1/16
0	0	0	1	0	0	1	1/16
0	0	1	0	0	1	1	1/16
0	0	1	1	0	1	2	1/16
0	1	0	0	1	1	1	1/16
0	1	0	1	1	1	2	1/16
0	1	1	0	1	2	2	1/16
0	1	1	1	1	2	3	1/16
1	0	0	0	1	1	1	1/16
1	0	0	1	1	1	2	1/16
1	0	1	0	1	2	2	1/16
1	0	1	1	1	2	3	1/16
1	1	0	0	0	0	0	1/16
1	1	0	1	0	0	1	1/16
1	1	1	0	0	1	1	1/16
1	1	1	1	0	1	2	1/16

Table 2: Indirect $XOR(X_1, X_2)$

Applying interventions, we would find that there is one direct link to Y and it is a link from X_3. As there is a direct link from X_3, by definition there is also a direct link from any set containing X_3, significance of these hyperlinks we evaluate in the information theoretical approach. If we further consider X_3 as target variable and X_1 and X_2 as source variables, there is also direct link from tuple (X_1, X_2) to X_3. If we compute all possible mutual and conditional mutual information (see table 3), we find out that the information-theoretical approach to direct causality is in agreement with the interventional one. Furthermore, as synergies $\text{Syn}(X_1, X_3)$ and $\text{Syn}(X_2, X_3)$ are equal zero, these hyperlinks are not considered to be uniquely causal. Synergy $\text{Syn}(X_1, X_2)$ is greater than zero but there is no direct link from tuple (X_1, X_2), therefore this hyperedge is also not included in hypergraph. Note that at this point we are focused just on dependencies of first or second order, otherwise we should investigate also hyperedge from all three variables. We will describe the concept of unique causality into more detail in the next section.

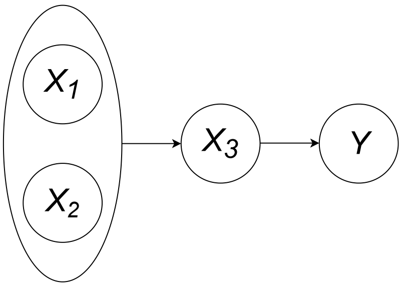


Figure 2: Causal hypergraph of system Indirect $XOR(X_1, X_2)$

3 Partial information decomposition

As we already said, if there is a direct link from one variable there is also a direct link from any set of variables including this variable. Therefore, in the causal hypergraph of the previous example (fig: 2.2) two hyperedges are missing. We decided to not draw these hyperedges because we do not consider them as unique. In this section, we introduce the concept of unique causality - one of the possibilities of non-negative decomposition of multivariate mutual information i.e. mutual information between a set of source variables and one target variable, suggested by Williams and Beer in [1]. In simplified form for two source variables, mutual information $I((X_1, X_2), Y)$ is decomposed into the sum of four functionals - unique contributions of X_1 and X_2, redundancy of these two variables and their synergy.

$$I((X_1, X_2), Y) = \mathrm{Un}(X_1) + \mathrm{Un}(X_2) + \mathrm{Red}(X_1, X_2) + \mathrm{Syn}(X_1, X_2) \tag{19}$$

The idea of individual functionals is as follows: X_1 may provide information that X_2 does not, this part of overall information we call unique information of X_1 about Y and denote as $\mathrm{Un}(X_1)$, unique information of X_2 about Y is defined analogously. Then, X_1 and X_2 may provide the same or overlapping information about Y, this overlapping part we call redundancy $\mathrm{Red}(X_1, X_2)$, for example if X_1 is a copy of X_2, they both provide same information about Y and thus $I((X_1, X_2), Y) = \mathrm{Red}(X_1, X_2)$ in this case. The last component is the so-called synergy, as we have seen in the XOR example, in some systems individual variables do not give us any information but as a tuple they do, $\mathrm{Syn}(X_1, X_2)$ should quantify this part of information about target Y. Since mutual information $I(X_1, Y)$ is equal to sum of $\mathrm{Un}(X_1)$ and $\mathrm{Red}(X_1, X_2)$, we also have

$$I((X_1, X_2), Y) = I(X_1, Y) + I(X_2, Y) - \mathrm{Red}(X_1, X_2) + \mathrm{Syn}(X_1, X_2). \tag{20}$$

First, mutual information between set of source variables $\mathbf{X} = (X_1, \ldots, X_n)$ and target variable Y can be express as

$$I(\mathbf{X}, Y) = \sum_{y \in \mathcal{Y}} p(y) I(\mathbf{X}, Y = y), \tag{21}$$

where $I(\mathbf{X}, Y = y)$ we call specific information of $\mathbf{X}$ about Y and it is defined as

$$I(\mathbf{X}, Y = y) = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}|y) \log \frac{p(y|\mathbf{x})}{p(y)}. \tag{22}$$

Information	bits
$I(X_1, Y)$	0.0000
$I(X_2, Y)$	0.0000
$I(X_3, Y)$	0.8113
$I(X_1, Y\|X_2)$	0.3113
$I(X_1, Y\|X_3)$	0.0000
$I(X_2, Y\|X_1)$	0.3113
$I(X_2, Y\|X_3)$	0.0000
$I(X_3, Y\|X_1)$	0.8113
$I(X_3, Y\|X_2)$	0.8113
$I(X_1, Y\|X_2, X_3)$	0.0000
$I(X_2, Y\|X_1, X_3)$	0.0000
$I(X_3, Y\|X_1, X_2)$	0.5000
$I((X_1, X_2), Y)$	0.3113
$I((X_1, X_3), Y)$	0.8113
$I((X_2, X_3), Y)$	0.8113
$I((X_1, X_2), Y\|X_3)$	0.0000
$I((X_1, X_3), Y\|X_2)$	0.8113
$I((X_2, X_3), Y\|X_1)$	0.8113
$I((X_1, X_2, X_3), Y)$	0.8113
$\mathrm{Syn}(X_1, X_2)$	0.3113
$\mathrm{Syn}(X_1, X_3)$	0.0000
$\mathrm{Syn}(X_2, X_3)$	0.0000

Table 3: Indirect $XOR(X_1X_2)$ - Information table

Idea how to define redundancy is then, that it is the expected value of the minimum information that any source variable provides about each outcome of Y. For two source variable X_1 and X_2 we get

$$\mathrm{Red}(X_1, X_2) = I_{\min}(Y, \{X_1\}\{X_2\}) = \sum_y p(y) \min_i I(Y = y, X_i). \tag{23}$$

Unique contributions are then defined as

$$\mathrm{Un}(X_1) = I(X_1, Y) - \mathrm{Red}(X_1, X_2) \tag{24}$$

$$\mathrm{Un}(X_2) = I(X_2, Y) - \mathrm{Red}(X_1, X_2) \tag{25}$$

and finally synergy of X_1 and X_2

$$\mathrm{Syn}(X_1, X_2) = I((X_1, X_2) - \mathrm{Un}(X_1) - \mathrm{Un}(X_2) - \mathrm{Red}(X_1, X_2). \tag{26}$$

As the synergies $\mathrm{Syn}(X_1, X_3)$ and $\mathrm{Syn}(X_2, X_3)$ are equal to zero we do not consider these hyperedges as valid.

4 Conclusion

In this work, we presented a general definition of the concept of causality designed by Judea Pearl and in practice often used information-theoretical approach. We discussed the problems with the information theoretical approach and definition of direct causality in systems with higher-order dependencies and came up with a new definition that seems consistent with the original definition, what we showed on selected examples. We also outline following algorithm for estimating causal hypergraphs of complex system from time series (or from data where we can sufficiently define target variable and source variables) which preserve directness and uniqueness of the hyperlinks: To detect parental hyperedges for individual target Y, we first find all variables with causal effect on Y using mutual information. For variables which did not have a causal effect on their own, we also compute multivariate mutual information between tuples of theese varaibles and the target variable, to find out if they have an effect as a pair (those that had an effect themselves force the effect as a tuple with any variable). Potentially continue with triplets and so on. Then determine the direct causality of each surviving variable and tuple. For each direct hyperedge, we then divide the causal effect to individual variable and tuple using partial information decomposition to get direct unique causal hypergraph. The algorithm in this form is however very inefficient, therefore, the next step of our research will be make this algorithm more efficient at least under some assumptions of the system. Also Beer and Williams approach to partial information decomposition can be replaced by another one.

References

[1] Williams, P.; Beer, R. Nonnegative Decomposition of Multivariate Information. *ArXiv* **2010**, abs/1004.2515.

[2] Pearl, J. Causality. Models, Reasoning, and Inference. 2nd ed. *Cambridge University Press* **2009**, doi: 10.1017/CBO9780511803161.

Non-augmented Neural Networks Robust to Rotation

Václav Košík

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

advisor: Jan Flusser, Department of Image Processing
Institute of Information Theory and Automation, CAS

Abstract. Convolutional neural networks are not invariant under basic image transformations like scaling, rotation or blur. Consequently, their performance falls drastically under influence of these transformations if they did not occur in the training set. The problem is usually tackled by augmenting the training set by these transformations. This, however, increases size of the training set multiple times and consequently the learning process is multiple times longer and in the case of massive datasets also much more expensive. We adjust architectures of CNNs to make their performance more robust to rotation. For that purpose, we make use of traditional and competitive approach to neural networks in image classification, so called handcrafted features. In particular, we use a handcrafted feature called Bispectrum which is invariant under rotation. The rotation was chosen as an easy case to work with. However, all the procedures are supposed to be generalized to other transformations in later research.

Keywords: Augmentation, Bispectrum, convolutional neural networks, invariants, rotation

Abstrakt. Konvoluční neuronové sítě nejsou invariantní na základní obrázkové transformace jako škálování, rotace nebo rozmazání obrázku. Jejich úspěšnost se proto velmi významně snižuje pod vlivem těchto transformací, pokud nefigurovaly v trénovací množině. Obvykle se tento problém řeší tzv. augmentováním trénovací množiny, což ale mnohonásobně zvyšuje její velikost, a tudíž i čas potřebný k trénování. V případě obrovských datasetů to i výrazně zvyšuje cenu trénování. V této práci přicházíme s novými architekturami konvolučních sítí, které jsou více robustní na rotaci obrázků. Využíváme k tomu tradičního a konkurenčního přístupu k neuronovým sítím v obrázkové klasifikaci, tzv. handcrafted příznaků. Konkrétně využíváme handcrafted příznak, který se nazývá Bispektrum a je invariantní na rotaci. Rotace byla zvolena jako jednoduchý případ deformace, ale v budoucím výzkumu se zaměříme na zobecnění všech postupů na jiné deformace.

Klíčová slova: Augmentace, Bispektrum, invarianty, konvoluční neuronové sítě, rotace

1 Introduction

Since early 1960's, researchers have been developing automatic image recognition and classification techniques. The traditional approach relies mainly on object description by means of features, which are measurable quantities that are able to uniquely characterize object classes. These kinds of features, currently denoted as "handcrafted", have been mostly designed either as results of local differential operators (such as SIFT and

SURF [11]) or as projections of the image function on a carefully selected functional basis (typical examples are harmonic basis leading to Fourier transform, various polynomial bases yielding image moments, wavelet basis and many others). These “projections” were used to define various sophisticated functionals, exhibiting the invariance with respect to intra-class variations and the ability to discriminate objects belonging to different classes. These functionals, called *invariants*, were then used as an input to a traditional classifier such as minimum-distance, SVM or Bayesian one [4]. For a comprehensive survey of handcrafted techniques see [5], Chapter 2, and further references therein.

Handcrafted features perform well in image recognition if there exist a (sufficiently simple) mathematical model of intra-class variations (typically if these variations comprise basic geometric transformations such as scaling, rotation, or global affine/projective transform for instance). In the case of generic classes with a wide intra-class variability, designing invariant and discriminative handcrafted features is usually very difficult or even impossible.

As an alternative to the handcrafted features, deep learning approach and convolutional neural networks (CNN) appeared in early 1980’s [6]. They were inspired by the visual perception in the human brain and went beyond the conventional framework which separates feature design from classifier training. However, during the first 30 years of their existence, they represented just a marginal research direction. They have attracted a noticeable attention of image processing community only since 2012, when a CNN named AlexNet won the ImageNet Large Scale Visual Recognition Challenge [17]. Soon the recognition rate have surpassed the human performance [8] thanks to a substantial increase of the computer performance at that time.

Instead of working with features "manufactured" beforehand, CNNs generate features by a cascade of convolutions and downsampling. Parameters of each convolution kernel are learned by a backpropagation algorithm. There are many convolution kernels in each layer, and each kernel is replicated over the entire image with the same parameters. The function of the convolution operators is to extract various features of the input. The network capacity depends on the number of layers. The first convolution layers obtain low-level features, such as edges, lines and corners. The more layers the network has, the higher-level features it produces. CNNs virtually skip the feature extraction step and require only basic preprocessing, which makes them, if enough training data and computing capacity are available, very powerful [2].

After a dynamic development in 2012-20, when many successful applications were reported, CNNs seem to have reached their limits. Further CNN’s development just by “evolution” is unlikely. Based on a comprehensive literature search and on our own experience, we have identified the following major drawbacks of current CNNs applied to image recognition.

- *Low-level image representation.* CNNs take a pixel-wise representation of images as an input, they do not perform any preprocessing, salient feature extraction, and other steps common in traditional image recognition. On the one hand, the pixel-level representation is highly redundant while on the other hand it is unable to capture even very simple intra-class variabilities.

- *Limited invariance.* Even if we have a model of intra-class variations, it is very difficult to incorporate the model into a CNN to achieve better efficiency. If the intra-class variability comprises rotation, scaling, and/or other simple-to-model transformations, the pixel-wise representation is not invariant to them. Current CNNs handle this by the *augmentation* of the training set, which is in fact a brute-force approach where we first artificially generate all possible transformations of training samples and the CNN is trained on this augmented set. Clearly, this is an extremely time and memory consuming process.

- *Massive training.* To achieve a good performance, CNNs should be trained on very large databases. This is partially implied by their limited invariance (see above) but even without the augmentation, the training set should be mostly much larger than in traditional approaches. Considering that the training set must be selected and annotated by domain experts, this is an expensive and time-consuming step.

- *The problem scale versus the computer performance.* In 2012-16, when CNNs exhibited a quick development and penetrated into many application areas, we witnessed a dynamic increase of computer performance. Powerful computers helped to resolve many large-scale problems, namely in image retrieval, despite the necessity of data augmentation and time-consuming training. However, nowadays the development seems to level out. Image databases collected by Google, Facebook and by many other commercial, research, and governmental organizations are so huge, that we need to make a qualitative step towards an efficient search and recognition. To rely just on continuous computer performance growth in a combination with current CNNs is not enough, because the problem scale increases faster than the computer performance.

CNNs have no true invariance hard coded by design. We have identified three approaches proposed in the literature that introduce invariance. The first one is similar to image normalization, e.g. in spatial transform networks [9], where a geometric transformation module is trained to put inputs/features into some predefined position while minimizing the network loss. Another approach is to transform inputs/features [10, 12] such that the intra-class variations appear as translations. The last approach transforms the convolution filters as e.g. in [15], Group Equivariant Convolution Networks [3] and Harmonic Networks [18]. All the methods can handle simple geometric transformations only. Generalization to more complex intra-class variations using these approaches is in most of the cases theoretically unfeasible.

The proposed methods of fusing handcrafted features with features learned by CNNs range from simple concatenation of features before classification or fusion at the score level [14, 13] to more advanced approaches such as learning features with handcrafted features as CNN's inputs [16, 1] or calculating handcrafted features from learned features [7]. However, none of the studies considered the possibility to harness invariant properties of handcrafted features.

In this paper, we incorporate a rotation invariant called Bispectrum to a CNN so it is more robust to a rotation transformation. Even very successful CNNs perform heavily worse on rotated images if the transformation was not included in the training set. This is usually handled by augmentation of the training set where each image is repeated several times with distinct rotations which increases the training set many times. Consequently, the training process is much longer. We are not able to make a CNN fully invariant to rotation without a bigger drop of performance on non-rotated images yet. However, the experiments show that we are at least able to significantly mitigate the degradation caused by this transformation without changing the training set at all. In our research, we will try to generalize all the results to other types of transformations (like scaling or blur) as well.

2 Bispectrum

A rotation in Cartesian coordinates is a shift in polar coordinates. Therefore, the idea of defining an invariant under rotation is to transform the image to polar coordinates because it is easier to get rid of a shift than a rotation. Let $I \in \mathbb{R}^{m,n}$ be an image function in polar coordinates with rows referring to $L2$-distance and columns referring to the angle of the image in Cartesian coordinates. If the original image is rotated by an angle α, the image in polar coordinates assigned to that rotated image is

$$I(r, \varphi + \alpha). \tag{1}$$

Bispectrum eliminates α in (1) and it does so for every row separately. Let $\mathcal{F}$ denote the Fourier transform. Then Bispectrum $B \in \mathbb{R}^{m,n}$ of an image $I \in \mathbb{R}^{m,n}$ in polar coordinates is defined as

$$B(k,\xi) = \mathcal{F}(I(k,\cdot))^2(\xi) \cdot \mathcal{F}^*(I(k,\cdot))(2\xi \mod n) \quad \text{for every } k \in \{1,2,\dots,m\}$$

where $\xi \in \{1,2,\dots,n\}$, all multiplication is element-wise and $*$ denotes complex conjugate.

Then k-th row of Bispectrum B_R of a rotated image $I(r,\varphi+\alpha)$ in polar coordinates is

$$\begin{aligned} B_R(k,\xi) &= \mathcal{F}(I(k,\cdot+\alpha))^2(\xi) \cdot \mathcal{F}^*(I(k,\cdot+\alpha))(2\xi \mod n) = \\ &= \mathcal{F}(I(k,\cdot))^2(\xi) \cdot \mathcal{F}^*(I(k,\cdot))(2\xi \mod n) e^{-2\pi i \xi 2\alpha} e^{2\pi i 2\xi\alpha} = \\ &= B(k,\xi). \end{aligned}$$

Therefore, Bispectrum is clearly invariant under rotation.

3 Methodology

3.1 Pixel model

First, we used a classic CNN with pixels of an image as an input. As we exhibited all the experiments on the MNIST dataset, we used a very simple architecture depicted on Fig. 1, i.e. two convolution layers (first one with 32 feature maps, the second one with 64) with max pooling, followed by a fully-connected layer. The second convolution layer was trained with 0.5 dropout regularization. The activation function is hyperbolic tangent on the convolution layers and softmax on the output layer which contains 10 nodes, each one corresponds to a class.

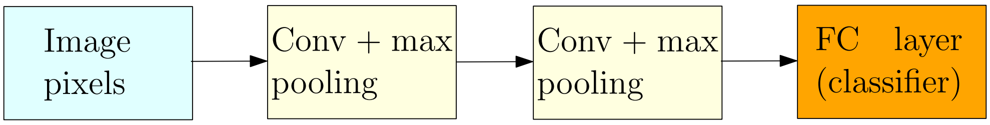


Figure 1: Architecture of the pixel model

3.2 Bispectral model

Second, we replaced the pixel-wise representation of an image by Bispectrum and used it as an input to a neural network. Such an architecture is obviously rotationally invariant. Moreover, Bispectrum contains full information of an image except for its rotation, i.e. there is an (computationally expensive) inversion of Bispectrum back to the image. The architecture remained the same as in the previous case, only the input was changed.

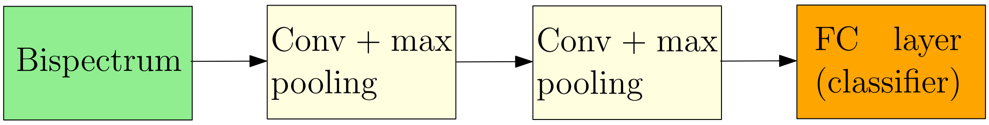


Figure 2: Architecture of the bispectral model

3.3 Parallel model

Third, we combined the first two approaches into one neural network, see Fig. 3. The network has two branches, the first branch takes pixels as an input, the second takes Bispectrum. Both branches have two convolutional layers with max pooling, the branches are connected together by a fully connected layer.

It turns out that if the network is learned like that, it considers almost only the pixel branch and assigns almost zero weights to the bispectral branch. We verified many times that the network has almost identical performance to the pixel model including the degradation when images are rotated. The reason is that automatic learning adjusts the weights only to the images which are in the training set. The optimization algorithm does not consider that the network will maybe once deal with rotated images where Bispectrum would be more relevant and important for better performance. The optimization is performed only for non-rotated images where pixels are more relevant and they can achieve excellent results on their own. Hence, the learning process basically switches off the second branch.

Therefore, the architecture must be adjusted to deal with this issue. We do that by pretraining the pixel and bispectral branch separately, i.e. by loading the convolutional weights from the first two models and freezing them in training. Hence, only the fully-connected layer is learned. Moreover, we change the loss function so that it penalizes the network for prioritizing the first branch over the second branch. Let W_1 denote the sum of all weights in absolute value going to the fully connected layer in the first branch. Let W_2 denote the same for the second branch. Then we adjust the classic sparse categorical

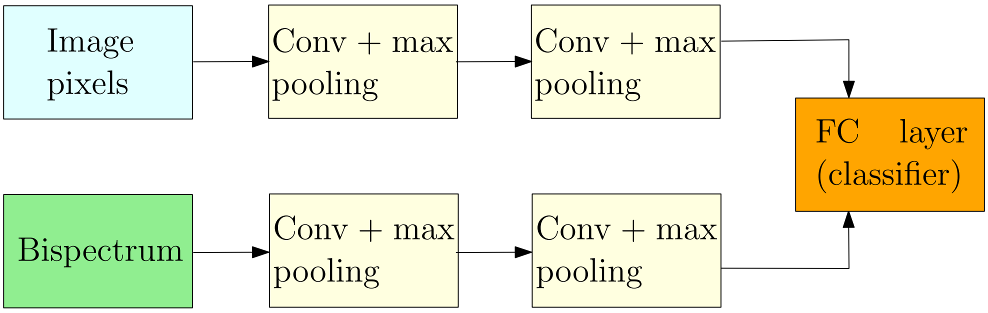


Figure 3: Architecture of the parallel model

crossentropy (which is used in all our other experiments) in the following way

$$\text{loss} = \text{sparse categorical crossentropy } + K \cdot \frac{W_1}{W_2} \tag{2}$$

where K is a constant. Therefore, if the network prioritizes the pixel branch over the bispectral branch, W_1 is greater than W_2 and the loss function grows.

3.4 Ensemble model

In the case of non-rotated images, the pixel model performs excellently. The output of the network are ten numbers, summed to 1, each signifies the probability that the image is certain digit. If the network accepts non-rotated image, one of these ten numbers is usually almost 1 while the others are almost zero because the network performs very good on MNIST. However, if we rotate an image at the beginning, the network is not that certain at all. Therefore, the principle of the ensemble is to let the pixel model compute and if the maximum of the ten numbers exceeds certain threshold, we believe that we deal with a non-rotated image and we let the pixel model decide. However, if the maximum is below that threshold, we expect it to be a rotated image and bispectral model decides. The threshold was set to 0.995.

4 Experiments

All experiments were performed on MNIST dataset of handwritten digits. The images are black and white with 28 rows and 28 columns. All models were trained on 60 000 samples without any augmentation, i.e. none of the images was rotated. We made three types of test set. The first one contains 10 000 non-rotated images, it is an original part of MNIST dataset. The second one contains the same images, but each one was randomly rotated by 30–90 degrees. The third set is union of the first and second one. Therefore, accuracy of the third set is arithmetic mean of the first two.

The K constant from the Equation (2) was chosen as 20. The optimization was performed by Adam algorithm with learning rate 0.001 in all cases. The batch size was set to 10.

To get Bispectrum, the image has to be first transformed to polar coordinates. It was done so by using incircle in the image. The images in polar coordinates are then of size $(14, 56)$ where the rows relate to $L2$ distance and columns to the angle. We chose 56 columns to keep the number of pixels the same. The linear interpolation was used. Bispectrum is then of size $(14, 56)$.

The results are summed in Table 1, the numbers represent accuracy. On the set of non-rotated images, the pixel model is the best with 99.28 % which is no surprise, even very simple CNNs perform excellently on MNIST dataset. However, we can see that the performance dropped drastically if we rotate the images. The bispectral model classified 82.63 % of images correctly. The little drop of performance in case of rotated images is probably caused by using incircle instead of excircle in the transformation to polar coordinates, so a bit of information can be lost. However, the model is almost invariant and performs much better in the case of rotated images. Next, parallel model and ensemble model are worse by only 2 % on non-rotated set then the pixel model, but both perform significantly better on rotated images, however worse then bispectral model. If both rotated and non-rotated images are equally presented, these two models outperform the others.

	Set 1	Set 2	Set 3
Pixel model	0.9928	0.4429	0.7179
Bispectral model	0.8263	0.8175	0.8219
Parallel model	0.9724	0.7691	0.8708
Ensemble model	0.9735	0.7981	0.8858

Table 1: Accuracies of our models on test sets of MNIST dataset. Set 1 is the original test set of MNIST containing non-rotated images. Set 2 contains the same images as Set 1 but randomly rotated by 30-90 degrees. Set 3 is union of Set 1 and Set 2.

In Table 2, we present the same results, but on images from the training set. These numbers are not so much different from the test sets.

	Set 1	Set 2	Set 3
Pixel model	0.9977	0.4526	0.7251
Bispectral model	0.8342	0.8282	0.8312
Parallel model	0.9766	0.7738	0.8752
Ensemble model	0.9775	0.8106	0.8941

Table 2: Accuracies of our models on training sets of MNIST dataset. Set 1 is the original training set of MNIST containing non-rotated images. Set 2 contains the same images as Set 1 but randomly rotated by 30-90 degrees. Set 3 is union of Set 1 and Set 2.

5 Discussion

We showed that the classic CNN performs very poorly on rotated images if the training set was not augmented by this rotation. That was an expected behaviour. If we replace the input by Bispectrum, we get a network invariant under rotation, but the accuracy drops to 83 %. The parallel and ensemble models have significantly smaller drop on non-rotated images and degrade much less on rotated images.

Although the ensemble gives the best results on the third set, its usage in other datasets might be questionable. MNIST dataset is very simple and even a CNN with only three layers can achieve almost perfect results. Therefore, we rely on the fact that the CNN classifies with very high certainty which does not have to be true in more complicated datasets.

Since Bispectrum contains full information about an image, it could theoretically get to the same numbers as the pixel-wise representation on non-rotated images. Theoretical justification for convolutions on pixels does not hold for Bispectrum which is in spectral domain. This could be a reason for the worse numbers, probably a different architecture would be more suitable. Moreover, if we make a small change in the pixel-wise representation, it can change Bispectrum significantly. This could be another reason for worse performance.

The rotation is only an example of a transformation which degrades CNN's performance if it is not included in the training set. We consider it as rather a simple example that helps us researching new architectures of neural networks. Then we want to explore more complicated transformations like blur and others and utilize the results and experience gained with rotation and invariants in general.

References

[1] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen. *Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification.* ISPRS Journal of Photogrammetry and Remote Sensing **138** (4 2018), 74–85.

[2] D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In 'Proceedings of the 22nd International Joint Conference on Artificial Intelligence IJCAI'11', volume 2, 1237–1242. AAAI Press, (2011).

[3] T. S. Cohen and M. Welling. Group equivariant convolutional networks. In 'Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML'16)', volume 48, 2990–2999, (2 2016).

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley Interscience, New York, USA, 2nd edition, (2001).

[5] J. Flusser, T. Suk, and B. Zitová. *2D and 3D Image Analysis by Moments.* Wiley, Chichester, U.K., (2016).

[6] K. Fukushima. *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position.* Biological Cybernetics **36** (1980), 193–202.

[7] Y. Hao, Q. Li, H. Mo, H. Zhang, and H. Li. *AMI-Net: Convolution neural networks with affine moment invariants.* IEEE Signal Processing Letters **25** (7 2018), 1064–1068.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In '2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', 770–778, (6 2016).

[9] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In 'Advances in Neural Information Processing Systems', volume 28, 2017–2025. Curran Associates, Inc., (2015).

[10] R. Jiang and S. Mei. Polar coordinate convolutional neural network: From rotation-invariance to translation-invariance. In '2019 IEEE International Conference on Image Processing (ICIP)', 355–359, (9 2019).

[11] L. Juan and O. Gwun. *A comparison of SIFT, PCA-SIFT and SURF.* International Journal of Image Processing (IJIP) **3** (2009), 143–152.

[12] J. Kim, W. Jung, H. Kim, and J. Lee. *CyCNN: A rotation invariant CNN using polar mapping and cylindrical convolution layers.* CoRR **abs/2007.10588** (2020).

[13] L. Nanni, S. Ghidoni, and S. Brahnam. *Handcrafted vs. non-handcrafted features for computer vision classification.* Pattern Recognition **71** (11 2017), 158–172.

[14] D. T. Nguyen, T. D. Pham, N. R. Baek, and K. R. Park. *Combining deep and hand-crafted image features for presentation attack detection in face recognition systems using visible-light camera sensors.* Sensors **18** (2 2018), 699.

[15] U. Schmidt and S. Roth. Learning rotation-aware features: From invariant priors to equivariant descriptors. In '2012 IEEE Conference on Computer Vision and Pattern Recognition', 2050–2057, (2012).

[16] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In 'Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)', volume 1, 568–576, (12 2014).

[17] Stanford Vision Lab. Imagenet large scale visual recognition challenge (ILSVRC), (2015). `http://www.image-net.org/challenges/LSVRC/`.

[18] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In '2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', 7168–7177, (2017).

Liouville-Green Approximation for Linearly Coupled Systems: Asymptotic Analysis with Applications to Reaction-Diffusion Systems

Juraj Kováč

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

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

Abstract. Multiple generalizations and modifications have been made to the Turing model for pattern formation in order to obtain a more faithful portrayal of the biological or ecological reality, most of which present the underlying reaction-diffusion (RD) equations with additional complexity. Among these generalizations is the concept of spatial heterogeneity which, mathematically speaking, turns the autonomous linearized RD system of ODEs into a non-autonomous one, making the search for an analytic solution all but futile. However, approximate methods can be used to investigate qualitative and even quantitative properties of the (usually) unknown exact solution, as has been the case with the WKBJ analysis presented by Krause et al.[1] In our present paper, we focus on building the general mathematical fundaments of such an analysis, providing approximation theorems for the Liouville-Green approximation (ı.e. WKBJ approximation in the absence of turning points) of the solution to linear systems of ODEs in one spatial dimension, including upper bounds of the error of such an approximation. We proceed by classifying systems by their spectral properties, carefully distinguishing between the exponential and the oscillatory cases, before merging the results into a single approximation theorem. It is worth noting, however, that our approach does not explicitly employ the usual WKBJ modes but rather exploits the well-known properties of the Airy functions, which display identical asymptotic behaviour, a fact readily used in the proof. Subsequently, we focus specifically on the RD equations, demonstrating the spectral properties utilized in the approximation theorems for a typical Turing system, hence arguing that the deployment of this asymptotic analysis is reasonable in the context of RD equations and Turing instability. As noted in the discussion, the arguably biggest shortcoming of our analysis is the exclusion of turning points and the resulting absence of connection formulae.

Keywords: Liouville-Green approximation, WKBJ, reaction-diffusion systems, Airy functions

Abstrakt. V snahe o vernejší popis biologickej či ekologickej reality sa Turingov model stal objektom viacerých modifikácií a zovšeobecnení, ktoré spravidla ďalej pridajú príslušnému systému reakčno-difúznych (RD) rovníc na komplexnosti. Medzi tieto zovšeobecnenia patrí aj koncept priestorovej heterogenity, ktorý - v reči matematiky - zmení linearizovaný RD systém ODR z autonómneho na neautonómny, čím sa nájdenie analytického riešenia (i linearizovanej sústavy) stáva prakticky nemožným. Použitie aproximatívnych metód však umožňuje skúmať nielen kvalitatívne, ale aj kvantitatívne vlastnosti tohto spravidla neznámeho presného riešenia, ako to ukázala analýza v podaní Krauseho et al. [1] Náš text sa sústredí na položenie matematických základov takejto asymptotickej analýzy prostredníctvom aproximačných teorémov pre Liouville-Greenovu aproximáciu (t.j. WKBJ aproximáciu bez bodov obratu) riešenia lineárnych systémov ODR v jednej priestorovej dimenzii, vrátane odhadov chyby takejto aproximácie. Náš postup klasifikuje systémy podľa spektrálnych vlastností, čo nám umožňuje dôsledne rozlišovať prípad exponenciálneho a oscilujúceho riešenia, aby sme oba výsledky následne zhrnuli do spoločného aproximačného teorému. Za zmienku azda stojí, že pri tom nevyužívame WKBJ módy v klasickom tvare, ale miesto toho užívame známych vlastností Airyho funkcií, ktoré vykazujú identické asymptotické chovanie, čo sa odzrkadľuje aj v príslušnom dôkaze. Následne svoju pozornosť obraciame priamo k RD rovniciam, u ktorých za 'turingovských' predpokladov nachádzame práve tie spektrálne vlastnosti, na ktorých je postavený dôkaz predchádzajúcich aproximačných teorémov. Tým demonštrujeme oprávnenosť využitia týchto asymptotických nástrojov na skúmanie RD rovníc a turingovskej nestability. Ako upozorňujeme v sekcii venovanej diskusii, najvýznamnejším nedostatkom našej analýzy je neuvažovanie bodov obratu a s tým súvisiaca absencia spojovacích formúl.

Kľúčové slová: Liouville-Greenova aproximácia, WKBJ, reakčno-difúzne systémy, Airyho funkcie

Full paper: J. Kováč and V. Klika. Liouville-Green approximation for linearly coupled systems: Asymptotic analysis with applications to reaction-diffusion systems. Discrete and Continuous Dynamical Systems – S 15 (2022), 2553–2579.

References

[1] A. Krause, V. Klika, T. Woolley, and E. Gaffney. *From one pattern into another: analysis of Turing patterns in heterogeneous domains via WKBJ.* Journal of The Royal Society Interface **17** (01 2020), 20190621.

New Classes of Quadratically Integrable Systems in Magnetic Fields: The Generalized Cylindrical and Spherical Cases[*]

Ondřej Kubů

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

advisors:
Libor Šnobl, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Antonella Marchesiello, Department of Applied Mathematics
Faculty of Information Technology, CTU in Prague

Abstract. We study integrable and superintegrable systems with magnetic field possessing quadratic integrals of motion on the three-dimensional Euclidean space. In contrast with the case without vector potential, the corresponding integrals may no longer be connected to separation of variables in the Hamilton–Jacobi equation and can therefore have more general leading order terms.

We focus on two cases extending the physically relevant cylindrical– and spherical–type integrals. We find three new integrable systems in the generalized cylindrical case but none in the spherical one. We conjecture that this is related to the presence, respectively absence, of maximal abelian Lie subalgebra of the three-dimensional Euclidean algebra generated by first order integrals in the limit of vanishing magnetic fields.

We find only one (minimally) superintegrable system among the integrable ones. It is the first system with a magnetic field which does not separate in any coordinate system. The results can be applied to the relativistic case as well due to vanishing scalar potential. This is crucial for the potential applications: Our system models an electron injected into a helical undulator inside an infinite solenoid, a key component of free electron lasers producing powerful pulses of circularly polarized radiation.

Keywords: integrability, superintegrability, magnetic field, generalized cylindrical and spherical cases, classical mechanics

Abstrakt. V této práci se zabýváme integrabilními a superintegrabilními systémy s magnetickým pole ve třídimenzionálním Euklidovském prostoru, které mají kvadratické integrály pohybu. Protože tyto integrály v případě s magnetickým polem nesouvisí se separací proměnných Hamilton-Jacobiho rovnice, mohou mít jejich členy nejvyššího řádu obecnější tvar.

Zaměřujeme se na dva fyzikálně relevantní případy, konkrétně rozšířený cylindrický a sférický. V rozšířeném cylindrickém případě nacházíme tři integrabilní systémy, v rozříříeném sférickém však žádný. Domníváme se, že by to mohlo souviset s maximální abelovskou Lieovou podalgebrou třídimenzionální Euklidovské algebry generovanou integrály prvního řádu, která v limitě

[*] This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS22/178/OHK4/3T/14

nulového magnetického pole buď je, nebo není přítomna.

Mezi těmito integrabilními systémy nacházime pouze jeden (minimálně) superintegrabilní systém. Jedná se o první systém s magnetickým polem, který není separabilní v žádné souřadné soustavě. Protože má tento systém nulový skalární potenciál, platí tyto výsledky i pro relativistický případ. To je zásadní pro potenciální aplikace, neboť náš systém popisuje elektron prolétavající šroubovicovým undulátorem, který je vnořen do solenoidu. Ten je klíčovou komponentou laseru na volných elektronech, který produkuje silné pulsy kruhově polarizovaného záření.

Klíčová slova: integrabilita, superintegrabilita, magnetické pole, zobecněný cylindrický a sférický případ, klasická mechanika

Full paper: O. Kubů, A. Marchesiello, L. Šnobl. *New classes of quadratically integrable systems in magnetic fields: the generalized cylindrical and spherical cases.* arXiv preprint, arXiv:2206.15305 (2022). Currently submitted for publication, undergoing peer review.

A Fibonacci's Complement Numeration System

Jana Lepšová

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

advisors:
Ľubomíra Dvořáková, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Sébastien Labbé, Laboratoire Bordelais de Recherche en Informatique
Université de Bordeaux & CNRS

Abstract. With the two's complement notation of signed integers, the fundamental arithmetic operations of addition, subtraction, and multiplication are identical to those for unsigned binary numbers. In this work, we consider a Fibonacci-equivalent of the two's complement notation. A transducer provided by Berstel computes the sum of the Zeckendorf binary representation of two nonnegative integers. In this work, we consider a numeration system also based on Fibonacci numbers but representing all integers. As for the two's complement notation, we show that addition of integers represented in this numeration system can be computed with the Berstel transducer with three additional transitions. Whether this can be done more generally is an open question raised by the current work.

Keywords: two's complement, numeration system, transducer, Fibonacci

Abstrakt. Two's complement (komplementární do dvou) reprezentace celých čísel mají vlastnost, že základní aritmetické operace sčítání, odečítání a násobení jsou stejné jako u klasických binárních reprezentací přirozených čísel. V tomto příspěvku představujeme analogii komplementárních reprezentací k Fibonacciho reprezentacím. Berstelův transducer z roku 1986 provádí sčítání Fibonacciho reprezentací dvou přirozených čísel. V tomto příspěvku zkoumáme numerační systém, který má jako bázi také Fibonacciho čísla, ale reprezentuje všechna celá čísla. Dokážeme, že sčítání dvou reprezentací celých čísel v tomto numeračním systému lze provést Berstelovým transducerem se třemi přidanými hranami. Součástí dalšího výzkumu je otázka, pro jaké další numerační systémy lze tento postup zobecnit.

Klíčová slova: two's complement, numerační systém, transducer, Fibonacci

1 Introduction

A nonnegative integer can be written as a sum of powers of 2 which gives rise to its binary expansion over alphabet $\Sigma = \{\mathtt{0}, \mathtt{1}\}$. Binary representations can be added with a standard algorithm - starting from the least significant digit and transferring a carry at each step. In the case that one of the representations is shorter in length, it is padded with the prefix of leading zeroes, as in the following example.

$$\begin{array}{rr} 11 & \texttt{01011} \\ +17 & \texttt{10001} \\ \hline 28 & \texttt{11100} \end{array} \qquad \text{(sum of binary representations)}$$

Among all the ways to generalize this approach to all integers including negative ones is the two's complement notation, see [11, §4.1]. In the two's complement representation of integers, the value of a binary word $w = w_k w_{k-1} \cdots w_1 w_0 \in \Sigma^{k+1}$ is

$$\mathrm{val}_{2c}(w) = \sum_{i=0}^{k-1} w_i 2^i - w_k 2^k. \tag{1}$$

It can be seen that for every $w \in \Sigma^*$, $\mathrm{val}_{2c}(\texttt{00}w) = \mathrm{val}_{2c}(\texttt{0}w)$ and $\mathrm{val}_{2c}(\texttt{11}w) = \mathrm{val}_{2c}(\texttt{1}w)$ and for every $n \in \mathbb{Z}$ there exists a unique word $w \in \Sigma^+ \setminus (\texttt{00}\Sigma^* \cup \texttt{11}\Sigma^*)$ such that $n = \mathrm{val}_{2c}(w)$. The word w is called the *two's complement representation* of the integer n, and we denote it by $\mathrm{rep}_{2c}(n)$.

The main interest with the two's complement notation is that the fundamental arithmetic operations of addition, subtraction, and multiplication are identical to those for unsigned binary numbers. For example, we perform below the addition of the representations seen previously, this time interpreting them in the two's complement notation. The first word has the same value $\mathrm{val}_{2c}(\texttt{01011}) = 2^3 + 2^1 + 2^0 = 11$ but this time $\mathrm{val}_{2c}(\texttt{10001}) = -2^4 + 2^0 = -15$.

$$\begin{array}{rr} 11 & \texttt{01011} \\ -15 & \texttt{10001} \\ \hline -4 & \texttt{11100} \end{array} \qquad \text{(sum of two's complement representations)}$$

The value of the resulting word is $\mathrm{val}_{2c}(\texttt{11100}) = -2^4 + 2^3 + 2^2 = -4$ which confirms the computation is correct. Notice that the negative integer -4 has a shorter two's complement representation and in particular $\mathrm{rep}_{2c}(-4) = \texttt{100}$.

Integers can also be expressed in other numeration systems [9, 8]. A typical example uses the Fibonacci numbers instead of the powers of 2. Let $(F_n)_{n \geq 0}$ be the Fibonacci sequence defined with the recurrence relation $F_n = F_{n-1} + F_{n-2}$, for all $n \geq 2$, and the initial conditions $F_0 = 1$, $F_1 = 2$, following a convention for the Fibonacci numeration system [5]. A result attributed to Zeckendorf [13, 4, 3] and published by Zeckendorf much later [20] (see also [10, Exercise 1.2.8.34]) says that every nonnegative integer n can be represented as a unique sum $n = \sum_{i=0}^{k} w_i F_i$ of nonconsecutive distinct Fibonacci numbers where $k = \max\{i \in \mathbb{N} : F_i \leq n\}$ and $w = w_k \cdots w_0 \in \Sigma^* \setminus \Sigma^*\texttt{11}\Sigma^*$. We refer to this numeration system on $\mathbb{N}$ as the Zeckendorf numeration system, we denote $\mathrm{val}_{\mathcal{Z}} : \Sigma^* \to \mathbb{N}$ its numerical value function and $\mathrm{rep}_{\mathcal{Z}} : \mathbb{N} \to \Sigma^* \setminus \Sigma^*\texttt{11}\Sigma^*$ its representation function.

In [1], an algorithm is given to compute the addition of nonnegative integers represented in the Zeckendorf numeration system. The representations $v, w \in \Sigma^k$ (of which the shorter in length is padded with leading zeros) are added digit by digit to obtain a word $u \in \{\texttt{0}, \texttt{1}, \texttt{2}\}^k$ of the same length $u = (v_{k-1} + w_{k-1}) \ldots (v_0 + w_0)$. The word u is given as input to a 10-state transducer $\mathcal{T}_\mathcal{Z}$ [1, p. 22] called *the adder* by Berstel reading from left to right. The word $\mathcal{T}_\mathcal{Z}(u) \in \Sigma^{k+3}$ is a binary word written only with $\texttt{0}$ and $\texttt{1}$ with the correct Zeckendorf value $\mathrm{val}_\mathcal{Z}(\mathcal{T}_\mathcal{Z}(u)) = \mathrm{val}_\mathcal{Z}(v) + \mathrm{val}_\mathcal{Z}(w)$.

$$\begin{array}{rl} 18 & \texttt{0101000} \\ +28 & \texttt{1001010} \\ \hline 46 & \texttt{1102010} \to \text{Berstel adder} \to \texttt{0010010101} \end{array} \qquad \text{(sum of Zeckendorf representations)}$$

However, it is not necessarily in the normal form, that is, the outputted word $\mathcal{T}_\mathcal{Z}(u)$ of the Berstel adder may contain consecutive $\texttt{1}$'s. It is known [1, 17, 6] that no single right-to-left and no single left-to-right transducer can normalize the word $u \in \{\texttt{0},\texttt{1},\texttt{2}\}^k$.

Motivated by the study of aperiodic tiling of the plane by Wang tiles, a numeration system $\mathcal{F}$ representing all integers in $\mathbb{Z}$ in a unique way based on Fibonacci numbers was introduced recently [12]. The goal of this contribution is to prove that it is the Fibonacci-equivalent of the two's complement notation. More precisely, we prove herein that the numeration system $\mathcal{F}$ is such that addition is performed using Berstel's adder (with three additional transitions) regardless of the sign of the entries. The numeration system $\mathcal{F}$ is based on a value map $\mathrm{val}_\mathcal{F} : \Sigma(\Sigma\Sigma)^* \to \mathbb{Z}$ defined for every odd-length binary words $w = w_{2k}w_{2k-1}\cdots w_0 \in \Sigma^{2k+1}$ as

$$\mathrm{val}_\mathcal{F}(w) = \sum_{i=0}^{2k-1} w_i F_i - w_{2k}F_{2k-1} \tag{2}$$

which is an analog of (1) using Fibonacci numbers instead of powers of 2.

The numeration system $\mathcal{F}$ extends naturally to $\mathbb{Z}^2$ and, in [12], it was used together with a certain automaton $\mathcal{A}$ to describe a particular aperiodic Wang shift Ω.

In this contribution we prove the following result which extends the two's complement arithmetical properties with respect to addition to the numeration system $\mathcal{F}$. We refer the reader to Definition 2.7 for the formal definition of the sum of two representations $\mathrm{rep}_\mathcal{F}(n) + \mathrm{rep}_\mathcal{F}(m)$ which involves the padding of the eventual shorter word with an appropriate neutral prefix.

Theorem 1.1. *Let $\mathcal{T}_\mathcal{F}$ be the Berstel adder $\mathcal{T}_\mathcal{Z}$ to which three transitions $S \xrightarrow{0|\varepsilon} 000.0$, $S \xrightarrow{1|\varepsilon} 101.7$ and $S \xrightarrow{2|\varepsilon} 100.6$ were added from a new initial state S replacing the original initial state. The transducer $\mathcal{T}_\mathcal{F}$ is a map $\{\texttt{0},\texttt{1},\texttt{2}\}^* \to \{\texttt{0},\texttt{1}\}^*$ such that*

$$\mathrm{val}_\mathcal{F}(\mathcal{T}_\mathcal{F}(\mathrm{rep}_\mathcal{F}(n) + \mathrm{rep}_\mathcal{F}(m))) = n + m$$

for every $n, m \in \mathbb{Z}$.

For example, using the numeration system $\mathcal{F}$, we compute

$$\begin{array}{rl} 18 & \texttt{0101000} \\ +(-6) & \texttt{1001010} \\ \hline 12 & \texttt{1102010} \to \text{modified Berstel adder} \to \texttt{110010101} \equiv \texttt{10101} \end{array}$$

Our main result is based on the Berstel adder $\mathcal{T}_\mathbb{Z}$ introduced in [1] for the addition of nonnegative integers in $\mathbb{N}$. A proof that Berstel adder works was provided in [7, Corollary 4] based on the numeration system in the real base $\tau = \frac{1+\sqrt{5}}{2}$, see also [5, 6]. Another proof follows from [2, §2.3.2.3] where it is proved that normalization in the real base β can be done with a finite automaton when β is a Pisot number. Pisot numbers are Parry numbers. We extend numeration systems associated to a subset of Parry numbers called simple Parry numbers to $\mathbb{Z}$ in the hope that in the ongoing work we will prove that they all have the property of addition on $\mathbb{N}$ and $\mathbb{Z}$ being performed by the same algorithm.

2 A Fibonacci Numeration System for $\mathbb{Z}$

In this section, we recall the numeration system $\mathcal{F}$ introduced in [12] which is defined by the value map $\mathrm{val}_\mathcal{F} : \Sigma(\Sigma\Sigma)^* \to \mathbb{Z}$ given in Equation (2) where $\Sigma = \{\mathtt{0}, \mathtt{1}\}$. The first observation to make on this value map is given in the next lemma.

Lemma 2.1. *For every word $w \in \Sigma^*$ of even length, we have*

$$\mathrm{val}_\mathcal{F}(\mathtt{000}w) = \mathrm{val}_\mathcal{F}(\mathtt{0}w) \quad \textit{and} \quad \mathrm{val}_\mathcal{F}(\mathtt{101}w) = \mathrm{val}_\mathcal{F}(\mathtt{1}w).$$

Proof. Let $w = w_{2k-1} \cdots w_0 \in \Sigma^*$ be of even length. We have

$$\mathrm{val}_\mathcal{F}(\mathtt{101}w) = \sum_{i=0}^{2k-1} w_i F_i + F_{2k} - F_{2k+1} = \sum_{i=0}^{2k-1} w_i F_i - F_{2k-1} = \mathrm{val}_\mathcal{F}(\mathtt{1}w). \qquad \square$$

Thus $\mathtt{00}$ or $\mathtt{10}$ can be used to pad words without changing their value.

Definition 2.2 (Neutral prefix). *Let $w \in \Sigma^*$ be of odd length. We say that* $\mathtt{00}$ *(*$\mathtt{10}$ *resp.) is the* neutral prefix *of w if $w \in \mathtt{0}\Sigma^*$ (if $w \in \mathtt{1}\Sigma^*$ resp.). We denote it by p_w.*

The following Lemma is an easy exercise on Fibonacci recurrence. It allows to determine the sign of $\mathrm{val}_\mathcal{F}(w)$ based only on the first digit of w.

Lemma 2.3. *For every word $w \in \Sigma^* \setminus \Sigma^* 11 \Sigma^*$ of even length we have*

1. $0 \leq \mathrm{val}_\mathcal{F}(0w) < F_{2k}$,
2. $-F_{2k+1} \leq \mathrm{val}_\mathcal{F}(100w) < 0$.

The following Proposition was proved in [12].

Proposition 2.4. *For every $n \in \mathbb{Z}$ there exists a unique odd-length word*

$$w \in \Sigma(\Sigma\Sigma)^* \setminus (\Sigma^*\mathtt{11}\Sigma^* \cup \mathtt{000}\Sigma^* \cup \mathtt{101}\Sigma^*)$$

such that $n = \mathrm{val}_\mathcal{F}(w)$.

Definition 2.5 (Numeration system $\mathcal{F}$ for $\mathbb{Z}$). *For each $n \in \mathbb{Z}$, we denote by $\mathrm{rep}_\mathcal{F}(n)$ the unique word satisfying the proposition.*

The neutral prefix can be used to pad words so that they all have the same length.

Definition 2.6 (Pad function)**.** *Let $u, v \in \Sigma(\Sigma\Sigma)^*$. We define*

$$\operatorname{pad}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \operatorname{pad}_k(u) \\ \operatorname{pad}_k(v) \end{pmatrix}$$

where $k = \max\{|u|, |v|\}$ and $\operatorname{pad}_k(w) = {p_w}^{\frac{1}{2}(k-|w|)}w$ for every $w \in \{u, v\}$ where p_w is the neutral prefix of the word w.

The padding allows us to define the sum of words or to represent coordinates in $\mathbb{Z}^d$ in dimension $d \geq 1$. Here we consider the case $d = 2$.

Definition 2.7 (Sum of two words)**.** *Let $\Sigma = \{\mathtt{0}, \mathtt{1}\}$ and $u, v \in \Sigma^*$. Then we define* $\operatorname{sum} : \Sigma^* \times \Sigma^* \to \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}^*$ *as*

$$\operatorname{sum}(u, v) = (u_{k-1} + v_{k-1}) \cdot \cdots \cdot (u_0 + v_0) \quad \text{where} \begin{pmatrix} u_{k-1} \dots u_0 \\ v_{k-1} \dots v_0 \end{pmatrix} = \operatorname{pad}\begin{pmatrix} u \\ v \end{pmatrix}.$$

Definition 2.8 (Numeration system $\mathcal{F}$ for $\mathbb{Z}^2$)**.** *Let $\boldsymbol{n} = (n_1, n_2) \in \mathbb{Z}^2$. We define*

$$\operatorname{rep}_{\mathcal{F}}(\boldsymbol{n}) = \operatorname{pad}\begin{pmatrix} \operatorname{rep}_{\mathcal{F}}(n_1) \\ \operatorname{rep}_{\mathcal{F}}(n_2) \end{pmatrix}.$$

In what follows, we need the following relation between the numeration system $\mathcal{F}$ and the usual Zeckendorf numeration system.

Lemma 2.9. *Let $u \in \Sigma^{2k+1}$ for some $k \geq 0$. Then $\operatorname{val}_{\mathcal{F}}(u) = \operatorname{val}_{\mathcal{Z}}(u) - u_{2k}F_{2k+1}$.*

Proof. The observation follows from

$$\operatorname{val}_{\mathcal{Z}}(u) = u_{2k}F_{2k} + \sum_{i=0}^{2k-1} u_iF_i = u_{2k}F_{2k+1} - u_{2k}F_{2k-1} + \sum_{i=0}^{2k-1} u_iF_i = u_{2k}F_{2k+1} + \operatorname{val}_{\mathcal{F}}(u). \quad \square$$

3 Addition of Zeckendorf representations on $\mathbb{N}$

A sequential transducer $\mathcal{T}$ as defined in [18] is a septuple $\mathcal{T} = (Q, A, B^*, \delta, \eta, i, \phi)$ where (Q, A, δ, i, ϕ) is a deterministic automaton over A^* with the partial function $\delta : Q \times A \to Q$, the output function $\eta : Q \times A \to B^*$ and the final function $\phi : Q \to B^*$. We restrict ourselves to letter-to-letter transducers, i.e. $\eta : Q \times A \to B$. Reading a word $u = u_\ell \dots u_0 \in A^*$, the transducer $\mathcal{T}$ moves between states $q_k \in Q$, with $q_0 = i$ and $q_{k+1} = \delta(q_k, u_k)$, outputting sequentially one letter $w_k = \eta(q_k, u_k) \in B$ for each input letter $u_k \in A$. After reading the whole word u the deterministic automaton is in a state $q_{\ell+1} \in Q$ and the value $\phi(q_{\ell+1})$ is concatenated at the end of the output word $w_\ell \dots w_0$, see also [7]. We write $\mathcal{T}(u) = w_\ell \dots w_0\phi(q_{\ell+1})$. By writing $\mathcal{T}(q, u)$ we mean that the transducer starts in an initial state $q \in Q$, i.e. $\mathcal{T}(i, u) = \mathcal{T}(u)$.

The main result of this article is based on a transducer proposed by Berstel in [1, p. 22] which we call the Berstel adder, reproduced in Figure 1.

Theorem 3.1. *The Berstel adder $\mathcal{T}_{\mathcal{Z}}$ fulfills that for every input $u \in \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}^*$, it outputs a word $\mathcal{T}_{\mathcal{Z}}(u) \in \{\mathtt{0}, \mathtt{1}\}^*$ with same value for the Zeckendorf numeration system:*

$$\operatorname{val}_{\mathcal{Z}}(u) = \operatorname{val}_{\mathcal{Z}}(\mathcal{T}_{\mathcal{Z}}(u)).$$

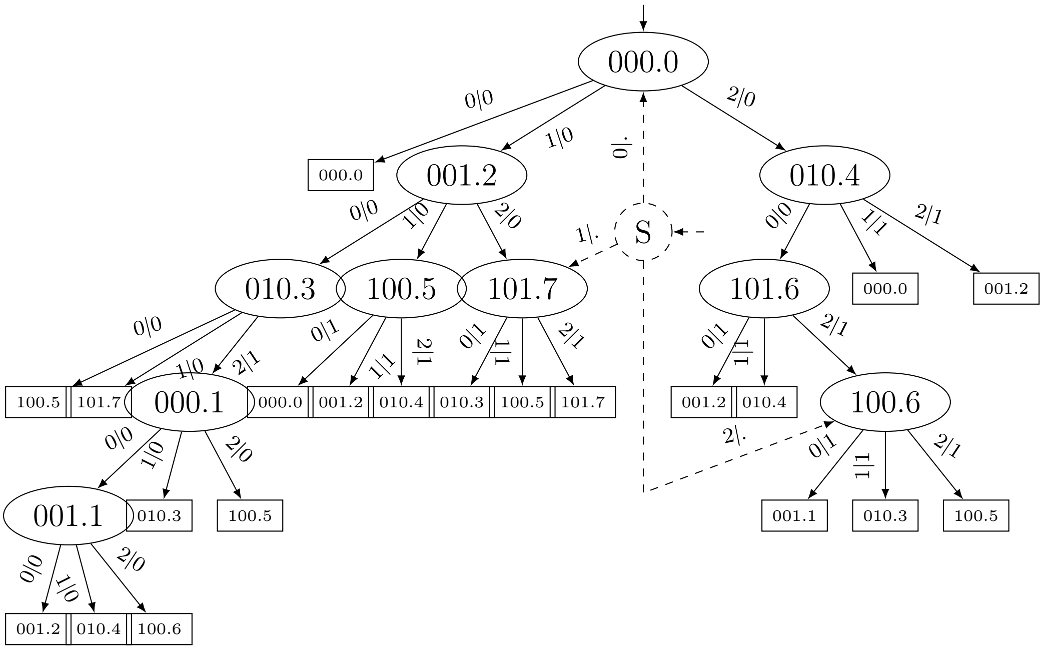


Figure 1: The solid edges represent the directed labeled graph G which which can be folded (after merging equivalent states) into the sequential transducer $\mathcal{T}_{\mathcal{Z}}$ known as the Berstel adder. A vertex reached from a path u has an ellipse shape if and only if u is minimal with respect to the radix order within the equivalence class $[u]_{\equiv}$ and has a rectangle shape if it $u \equiv v$ for some word $v <_{rad} u$. The solid and dashed edges with the additional initial state S represent the sequential transducer $\mathcal{T}_{\mathcal{F}}$ which deals with representations of negative integers as well.

4 Addition of representations in $\mathcal{F}$ on $\mathbb{Z}$

The algorithm for addition on $\mathbb{Z}$ takes two integers $n_1, n_2 \in \mathbb{Z}$ as input. The vector $\boldsymbol{n} = (n_1, n_2) \in \mathbb{Z}^2$ is represented in the numeration system $\mathcal{F}$. Then the coordinates of $\mathrm{rep}_{\mathcal{F}}(n_1, n_2)$ are are added digit by digit, giving rise to a word $u = \mathrm{sum}(\boldsymbol{n})$ on the alphabet $\{\mathtt{0}, \mathtt{1}, \mathtt{2}\}$. In this section, we prove Theorem 1.1.

We derive a transducer $\mathcal{T}_{\mathcal{F}}$ from $\mathcal{T}_{\mathcal{Z}}$ as $\mathcal{T}_{\mathcal{F}} = (Q \cup \{S\}, \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}, \{\mathtt{0}, \mathtt{1}\}, \delta_{\mathcal{F}}, \eta_{\mathcal{F}}, S, \phi)$ by adding a new initial state S and extending $\delta_{\mathcal{Z}}$ and $\eta_{\mathcal{Z}}$ of $\mathcal{T}_{\mathcal{Z}}$ in the following way

- for every $q \in Q$ and every $u_0 \in \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}$, $\delta_{\mathcal{F}}(q, u_0) = \delta_{\mathcal{Z}}(q, u_0)$ and $\eta_{\mathcal{F}}(q, u_0) = \eta_{\mathcal{Z}}(q, u_0)$,
- $\delta_{\mathcal{F}}(S, 0) = \mathtt{000.0}$, $\delta_{\mathcal{F}}(S, 1) = \mathtt{101.7}$, $\delta_{\mathcal{F}}(S, 2) = \mathtt{100.6}$,
- $\eta_{\mathcal{F}}(S, u_0) = \varepsilon$ for every $u_0 \in \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}$.

Obviously starting from every state $q \in Q$ different than S, the transducers $\mathcal{T}_{\mathcal{Z}}$ and $\mathcal{T}_{\mathcal{F}}$ have the same output for every $u \in \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}^*$, i.e. $\mathcal{T}_{\mathcal{Z}}(q, u) = \mathcal{T}_{\mathcal{F}}(q, u)$.

Proof of Theorem 1.1. I. Let $u = \mathtt{0}u' \in \mathtt{0}\{\mathtt{0}, \mathtt{1}, \mathtt{2}\}^{2k}$. We observe that from the definition of the transducers $\mathcal{T}_{\mathcal{F}}$ and $\mathcal{T}_{\mathcal{Z}}$ that $\mathcal{T}_{\mathcal{F}}(\mathtt{0}u') = \mathcal{T}_{\mathcal{Z}}(u')$ and $\mathcal{T}_{\mathcal{F}}(u) \in \mathtt{0}\Sigma^{2k+2}$. Then using these observations, Lemma 2.9 and Theorem 3.1 we derive

$$\mathrm{val}_{\mathcal{F}}(\mathcal{T}_{\mathcal{F}}(u)) = \mathrm{val}_{\mathcal{Z}}(\mathcal{T}_{\mathcal{F}}(u)) = \mathrm{val}_{\mathcal{Z}}(\mathcal{T}_{\mathcal{Z}}(u')) = \mathrm{val}_{\mathcal{Z}}(u') = \mathrm{val}_{\mathcal{Z}}(u) = \mathrm{val}_{\mathcal{F}}(u).$$

II. Let $u \in \mathtt{1}\{\mathtt{0},\mathtt{1},\mathtt{2}\}^{2k}$. Then either $u = \mathtt{10}u'$ or $u = \mathtt{11}u'$ for some $u' \in \Sigma^{2k-1}$. From the definition of the transducers $\mathcal{T}_{\mathcal{F}}$ and $\mathcal{T}_{\mathcal{Z}}$ we observe that $\delta_{\mathcal{F}}(S, \mathtt{10}) = \delta_{\mathcal{Z}}(\mathtt{000.0}, \mathtt{10})$ and $\delta_{\mathcal{F}}(S, \mathtt{11}) = \delta_{\mathcal{Z}}(\mathtt{000.0}, \mathtt{11})$. Moreover,

$$\begin{aligned} \mathcal{T}_{\mathcal{F}}(\mathtt{10}u') &= \mathtt{1}\mathcal{T}_{\mathcal{F}}(\mathtt{010.3}, u'), & \mathcal{T}_{\mathcal{Z}}(\mathtt{10}u') &= \mathtt{00}\mathcal{T}_{\mathcal{Z}}(\mathtt{010.3}, u') = \mathtt{00}\mathcal{T}_{\mathcal{F}}(\mathtt{010.3}, u'), \\ \mathcal{T}_{\mathcal{F}}(\mathtt{11}u') &= \mathtt{1}\mathcal{T}_{\mathcal{F}}(\mathtt{100.5}, u'), & \mathcal{T}_{\mathcal{Z}}(\mathtt{11}u') &= \mathtt{00}\mathcal{T}_{\mathcal{Z}}(\mathtt{100.5}, u') = \mathtt{00}\mathcal{T}_{\mathcal{F}}(\mathtt{100.5}, u'). \end{aligned}$$

Thus we can summarize that for $u \in \mathtt{1}\{\mathtt{0},\mathtt{1},\mathtt{2}\}^{2k}$ there exists a unique $w \in \Sigma^{2k+2}$ such that $\mathcal{T}_{\mathcal{F}}(u) = \mathtt{1}w$ and $\mathcal{T}_{\mathcal{Z}}(u) = \mathtt{00}w$. Using the previous observation, Lemma 2.9 and Theorem 3.1, we derive the statement

$$\begin{aligned} \mathrm{val}_{\mathcal{F}}(u) + F_{2k+1} &= \mathrm{val}_{\mathcal{Z}}(u) = \mathrm{val}_{\mathcal{Z}}(\mathcal{T}_{\mathcal{Z}}(u)) = \mathrm{val}_{\mathcal{Z}}(w) \\ &= \mathrm{val}_{\mathcal{Z}}(\mathtt{1}w) - F_{2k+2} = \mathrm{val}_{\mathcal{F}}(\mathtt{1}w) + F_{2k+3} - F_{2k+2} = \mathrm{val}_{\mathcal{F}}(\mathcal{T}_{\mathcal{F}}(u)) + F_{2k+1}. \end{aligned}$$

III. Let $u \in \mathtt{2}\{\mathtt{0},\mathtt{1},\mathtt{2}\}^{2k}$. Then either $u = \mathtt{200}u'$ or $u = \mathtt{201}u'$ for some $u' \in \Sigma^{2k-2}$. From the definition of the transducers $\mathcal{T}_{\mathcal{F}}$ and $\mathcal{T}_{\mathcal{Z}}$ we observe that $\delta_{\mathcal{F}}(S, \mathtt{200}) = \delta_{\mathcal{Z}}(\mathtt{000.0}, \mathtt{200})$ and $\delta_{\mathcal{F}}(S, \mathtt{201}) = \delta_{\mathcal{Z}}(\mathtt{000.0}, \mathtt{201})$. Moreover,

$$\begin{aligned} \mathcal{T}_{\mathcal{F}}(\mathtt{200}u') &= \mathtt{10}\mathcal{T}_{\mathcal{F}}(\mathtt{001.2}, u'), & \mathcal{T}_{\mathcal{Z}}(\mathtt{200}u') &= \mathtt{001}\mathcal{T}_{\mathcal{Z}}(\mathtt{001.2}, u') = \mathtt{001}\mathcal{T}_{\mathcal{F}}(\mathtt{001.2}, u'), \\ \mathcal{T}_{\mathcal{F}}(\mathtt{201}u') &= \mathtt{10}\mathcal{T}_{\mathcal{F}}(\mathtt{010.4}, u'), & \mathcal{T}_{\mathcal{Z}}(\mathtt{201}u') &= \mathtt{001}\mathcal{T}_{\mathcal{Z}}(\mathtt{010.4}, u') = \mathtt{001}\mathcal{T}_{\mathcal{F}}(\mathtt{010.4}, u'). \end{aligned}$$

Thus we can summarize that for $u \in \mathtt{2}\{\mathtt{0},\mathtt{1},\mathtt{2}\}^{2k}$ there exists a unique $w \in \Sigma^{2k+1}$ such that $\mathcal{T}_{\mathcal{F}}(u) = \mathtt{10}w$ and $\mathcal{T}_{\mathcal{Z}}(u) = \mathtt{001}w$. Using the previous observation, Lemma 2.9 and Theorem 3.1, we derive the statement

$$\begin{aligned} \mathrm{val}_{\mathcal{F}}(u) + 2F_{2k+1} &= \mathrm{val}_{\mathcal{Z}}(u) = \mathrm{val}_{\mathcal{Z}}(\mathcal{T}_{\mathcal{Z}}(u)) = \mathrm{val}_{\mathcal{Z}}(\mathtt{1}w) = \mathrm{val}_{\mathcal{Z}}(\mathtt{10}w) + F_{2k+1} - F_{2k+2} \\ &= \mathrm{val}_{\mathcal{Z}}(\mathtt{10}w) - F_{2k} = \mathrm{val}_{\mathcal{F}}(\mathtt{10}w) + F_{2k+3} - F_{2k} = \mathrm{val}_{\mathcal{F}}(\mathcal{T}_{\mathcal{F}}(u)) + 2F_{2k+1}, \end{aligned}$$

where we used the property that for every $\ell \geq 0$, $2F_{2\ell+1} = F_{2\ell+3} - F_{2\ell}$. □

5 Complement version of Simple Parry numeration system

In this section, we explore an extension of the numeration system $\mathcal{F}$ to numeration systems based on simple Parry numbers, see [8]. Fibonacci numbers are closely related to the golden mean $\tau = \frac{1+\sqrt{5}}{2}$ which is a simple Parry number.

Parry numbers are real numbers $\beta > 1$ with Rényi expansion of unity $d_\beta(1) = (t_i)_{i \geq 1}$ which is eventually periodic. The Parry numbers whose Rényi expansion of unity $d_\beta(1)$ is finite are called simple Parry numbers. The Rényi expansion of unity in the base β which is finite is such that $d_\beta(1) = t_1 \dots t_m$ implies $1 = \sum_{i=1}^{m} \frac{t_i}{\beta^i}$ (see [16] for the formal definition). The infinite Rényi expansion of unity $d^*_\beta(1)$ of a simple Parry number is then defined as $d^*_\beta(1) = \lim_{x \to 1^-} d_\beta(x) = (t_1 \dots t_{m-1}(t_m - 1))^\omega$.

For all real $\beta > 1$, if we denote $d^*_\beta(1) = (t_i)_{i\geq 1}$, there is a numeration system U_β canonically associated with β defined be

$$U_n = t_1 U_{n-1} + \cdots + t_n U_0 + 1, \text{ for all } n \geq 0. \tag{3}$$

We denote Σ_β the alphabet $\Sigma_\beta = \{0, 1, \ldots, C_U - 1\}$ where $C_U := \sup_{n\geq 0} \lceil U_{n+1}/U_n \rceil < +\infty$. The following steps may be done more generally for all Parry numbers but for simplicity we restrict ourselves to simple Parry numbers. For the canonical systems U_β associated to $\beta > 1$ simple Parry numbers with $d_\beta(1) = t_1 \ldots t_m$, we can derive a linear recurrence relation

$$U_n = \sum_{i=1}^{m} t_i U_{n-i} \text{ for every } n \geq m. \tag{4}$$

The golden mean τ fulfills that $d_\tau(1) = 11$ and we observe that $1 = \frac{1}{\tau} + \frac{1}{\tau^2}$. Thus the golden mean is indeed a simple Parry number and $d^*_\tau(1) = (10)^\omega$. The linear recurrence relation of the associated numeration system U_τ is exactly the Fibonacci recurrence.

For $\beta > 1$ a simple Parry number, the greedy representations $\mathrm{rep}_\beta(n)$ of nonnegative integers $n \in \mathbb{N}$ possibly preceded by leading zeroes form a regular language accepted by a deterministic finite automaton (DFA) denoted $\mathcal{A}_{\beta,q_1}$ (see [15]), in other words

$$\mathcal{L}(\mathcal{A}_{\beta,q_1}) = 0^* \,\mathrm{rep}_\beta(\mathbb{N}). \tag{5}$$

The DFA $\mathcal{A}_{\beta,q_1}$ has the set of states $Q_\beta = \{q_1, \ldots, q_m\}$, all of which are final. The initial state is q_1. For all $j \in \{1, \ldots, m\}$ there are t_j edges from q_j to q_1 labeled $0, \ldots, t_j - 1$ and for all $j \in \{1, \ldots, m-1\}$ there is one edge from q_j to q_{j+1} labeled t_j. We denote $\mathcal{L}_n(\mathcal{A}_{\beta,q_1}) = \mathcal{L}(\mathcal{A}_{\beta,q_1}) \cap \Sigma^n_\beta$.

Lemma 5.1. *The automaton $\mathcal{A}_{\beta,q_1}$ fulfils that $\#\mathcal{L}_n(\mathcal{A}_{\beta,q_1}) = U_n$ for every $n \in \mathbb{N}$.*

Let us denote $\mathcal{A}_{\beta,q_k}$ the DFA which arises from $\mathcal{A}_{\beta,q_1}$ by changing its initial state to q_k for $k \in \{1, \ldots, m\}$. With Lemma 5.1 in mind, we define $V_{n,k} = \#\mathcal{L}_n(\mathcal{A}_{\beta,q_k})$ for every $n \in \mathbb{Z}$. As a consequence of Lemma 5.1, $V_{n,1} = U_n$ for every $n \in \mathbb{N}$. Naturally, $V_{n,k} = 0$ for all $n < 0$. For every word $w = w_N w_{N-1} \ldots w_0 \in \Sigma^{N+1}_\beta$ for some $N \in \mathbb{N}$, we define

$$\mathrm{val}^k_\beta(w) = \textstyle\sum_{i=0}^{N-1} w_i U_i - w_N V_{N,k}. \tag{6}$$

We extend the automaton $\mathcal{A}_{\beta,q_k}$ to an automaton $\mathcal{A}^k_\beta$ by creating a new initial state S and adding two new edges $S \xrightarrow{0} q_1$ and $S \xrightarrow{1} q_k$ so that $\mathcal{L}(\mathcal{A}_\beta) = 0\mathcal{L}(\mathcal{A}_{\beta,q_1}) \cup 1\mathcal{L}(\mathcal{A}_{\beta,q_k})$.

Proposition 5.2. *Let $k \in \{1, \ldots, m\}$. For every $n \in \mathbb{Z}$ there exists a unique word $u \in \mathcal{L}(\mathcal{A}^k_\beta) \cap \Sigma_\beta(\Sigma^m_\beta)^* \setminus (00^m\Sigma^*_\beta \cup 1p_k\Sigma^*_\beta)$ such that $\mathrm{val}^k_\beta(u) = n$, where $p_k = t_k \ldots t_{m-1}(t_m - 1)t_1 \ldots t_{k-1}$ is the associated neutral padding word.*

Note that in the numeration system $\mathcal{F}$, the padding by neutral prefix $p_w = \mathtt{10}$ of a word $w = \mathtt{1}w' \in \mathtt{1}\Sigma^*$ can be seen in this more general setup as putting a neutral padding word $p_k = \mathtt{01}$ after the initial letter $\mathtt{1}$, i.e. $p^*_w\mathtt{1}w' = (\mathtt{10})^*\mathtt{1}w' = \mathtt{1}(\mathtt{01})^*w' = \mathtt{1}p^*_k w'$.

Definition 5.3 (Numeration system $V_{\beta,k}$)**.** *Let $k \in \{1, \ldots, m\}$. For $n \in \mathbb{Z}$ we denote by $\mathrm{rep}^k_\beta(n)$ the unique word u from Proposition 5.2.*

If U_β is a numeration system associated to β a Pisot number (Pisot numbers are Parry numbers, see [19]) such that the characteristic polynomial of U_β is the minimal polynomial of the number β, then addition of nonnegative integers in the numeration system U_β can be performed by a finite transducer, see [14, § 7].

Question 5.4. *Let $\beta > 1$ be a simple Parry number with its associated transducer for addition on $\mathbb{N}$. Can we extend this transducer to handle addition on $\mathbb{Z}$ using the complement version of the numeration system $V_{\beta,k}$?*

To achieve this, it is needed to extend Theorem 3.1.

6 Proofs of results in Section 5

Proof of Lemma 5.1. From Equation (5) we have $\mathcal{L}_n(\mathcal{A}_{\beta,q_1}) = \bigsqcup_{i=0}^{n}\{0^i \operatorname{rep}_\beta(j) : |\operatorname{rep}_\beta(j)| = n-i\}$. From the greediness of the representations rep_β we have that for any integer $n \in \mathbb{N}$, $\#\{i \in \mathbb{N} : |\operatorname{rep}_\beta(i)| = n\} = U_n - U_{n-1}$ (where we consider $U_{-1} = 0$). Together,

$$\#\mathcal{L}_n(\mathcal{A}_{\beta,q_1}) = \sum_{i=0}^{n}(U_i - U_{i-1}) = U_n \text{ for every } n \in \mathbb{N}. \qquad \square$$

We state the following observation without proof.

Remark 6.1. *For any word $w \in \Sigma_\beta^*$, if $w \in \mathcal{L}(\mathcal{A}_{\beta,q_k})$ then $w \in \mathcal{L}(\mathcal{A}_{\beta,q_1})$.*

Lemma 6.2. *$V_{n,k} = \sum_{d=k}^{m} t_d U_{n-(d-k+1)}$ for any $n \geq m-k+1$.*

Proof. The paths starting at q_k of length $n \in \mathbb{N}$ reach the state q_1 after at most $m-k+1$ edges. Let d denote the smallest number of edges after which a path of length $n \in \mathbb{N}$ in the DFA $\mathcal{A}_{\beta,q_k}$ reaches the state q_1. Then

$$\mathcal{L}_n(\mathcal{A}_{\beta,q_k}) = \bigsqcup_{d=1}^{m-k+1} \mathcal{L}_{n-d}(\mathcal{A}_{\beta,q_1}) = \bigsqcup_{d=k}^{m} \mathcal{L}_{n-(d-k+1)}(\mathcal{A}_{\beta,q_1}).$$

For any $i \in \{1,\ldots,k\}$, the amount of different paths from q_i to q_1 of minimal non-zero length is t_i. The result is a consequence of Lemma 5.1. $\square$

We omit in this material the technical proof of the following Lemma which is based on the reccurence formulas (3), (4) and Lemma 6.2.

Lemma 6.3. *Let $k \in \{0,\ldots,m\}$ and $\ell \in \mathbb{N}$. Then for any $w \in \mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_k})$ we have*

$$\operatorname{val}_\beta^k(1w) = \operatorname{val}_\beta^k(1t_kt_{k+1}\ldots t_{m-1}(t_m-1)t_1\ldots t_{k-1}w).$$

We denote $p_k = t_k \ldots t_{m-1}(t_m - 1)t_1 \ldots t_{k-1}$ the neutral padding word for a certain $k \in \{1,\ldots,m\}$. We observe that p_k is a cyclic permutation of the smallest period of $d_\beta^*(1)$ and at the same time a cycle in the automaton $\mathcal{A}_{\beta,q_k}$ following the maximal edges.

Lemma 6.4. *Let $k \in \{1,\ldots,m\}$ and $\ell \in \mathbb{N}$. The map $w \mapsto \operatorname{val}_\beta^k(1w)$ is a bijection from the set $\mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_k}) \setminus p_k\Sigma_\beta^*$ to the interval of integers $\{n \in \mathbb{Z} \mid -V_{m\ell,k} \leq n < -V_{m\ell-m,k}\}$.*

Proof. We see that if $|w| = 0$ then $-V_{0,k} = \mathrm{val}_\beta^k(1) - 1 < 0 = -V_{-m,k}$. Let us assume that $|w| = m\ell$ for $\ell \geq 1$. Then

$$\mathrm{val}_\beta^k(1w) \geq \mathrm{val}_\beta^k(10^{m\ell}) = -V_{m\ell,k}.$$

The lexicographically largest word $w \in \mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_k}) \setminus p_k\Sigma_\beta^*$ is

$$w_{\max} = t_k \dots t_{m-1}(t_m - 1)t_1 \dots t_{k-2}(t_{k-1} - 1)(t_1 \dots t_{m-1}(t_m - 1))^{\ell-1}.$$

Then we see that

$$\begin{aligned}\mathrm{val}_\beta^k(1w) \leq \mathrm{val}_\beta^k(1w_{\max}) &= \mathrm{val}_\beta^k(1p_k0^{(\ell-1)m}) - U_{(\ell-1)m} + \mathrm{val}_\beta((t_1 \dots t_{m-1}(t_m - 1))^{\ell-1}) \\ &= \mathrm{val}_\beta^k(1p_k0^{(\ell-1)m}) - U_{(\ell-1)m} + U_{(\ell-1)m} - 1 = -V_{m\ell-m} - 1 < -V_{m\ell-m}\end{aligned}$$

Therefore the codomain of the map $w \mapsto \mathrm{val}_\beta^k(1w)$ is indeed the interval of integers $\{n \in \mathbb{Z} \mid -V_{m\ell,k} \leq n < -V_{m\ell-m,k}\}$.

Next we show that the domain and the codomain of the map $w \mapsto \mathrm{val}_\beta^k(1w)$ have the same cardinality. Indeed, $\mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_k}) \setminus p_k\Sigma_\beta^* = \mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_k}) \setminus p_k\mathcal{L}_{m\ell-m}(\mathcal{A}_{\beta,q_k})$ and thus $\#\mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_k}) \setminus p_k\Sigma_\beta^* = V_{m\ell,k} - V_{m\ell-m,k} = \#\{n \in \mathbb{Z} \mid -V_{m\ell,k} \leq n < -V_{m\ell-m,k}\}$.

It suffices to show that the map $w \mapsto \mathrm{val}_\beta^k(1w)$ is injective. We assume by contradiction the existence of $w, w' \in \mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_k}) \setminus p_k\Sigma_\beta^*$ such that $\mathrm{val}_\beta^k(1w) = \mathrm{val}_\beta^k(1w')$. Consequently, $\mathrm{val}_\beta(w) = \mathrm{val}_\beta(w')$. Let $i, i' \in \mathbb{N}$ be the maximal exponents so that $w = 0^iv$ and $w' = 0^{i'}v'$ for some words $v, v' \in \Sigma_\beta^*$. By Remark 6.1, $v, v' \in \mathcal{L}(\mathcal{A}_{\beta,q_1})$ and v, v' are greedy representations in the numeration system U_β fulfilling $\mathrm{val}_\beta(v) = \mathrm{val}_\beta(v')$. It follows that $v = v'$ and thus $w = w'$. □

Lemma 6.5. *Let $k \in \{1, \dots, m\}$ and $\ell \in \mathbb{N}$. The map $w \mapsto \mathrm{val}_\beta^k(0w)$ is a bijection from the set $\mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_1}) \setminus 0^m\Sigma_\beta^*$ to the interval of integers $\{n \in \mathbb{N} \mid U_{m(\ell-1)} \leq n < U_{m\ell}\}$.*

Proof. We see that if $|w| = 0$ then $U_{-1} = \mathrm{val}_\beta^k(0)0 < 1 = U_0$. Let us assume that $|w| = m\ell$ for $\ell \geq 1$. Then

$$\mathrm{val}_\beta^k(0w) \geq \mathrm{val}_\beta^k(00^{m-1}10^{m(\ell-1)}) = U_{m(\ell-1)}.$$

On the other hand, the word w is a greedy representation possibly preceeded by leading zeroes and therefore

$$\mathrm{val}_\beta^k(0w) = \mathrm{val}_\beta w < U_{m\ell}.$$

Therefore the codomain of the map $w \mapsto \mathrm{val}_\beta^k(0w)$ is indeed the interval of integers $\{n \in \mathbb{N} \mid U_{m(\ell-1)} \leq n < U_{m\ell}\}$.

Next we show that the domain and the codomain of the map $w \mapsto \mathrm{val}_\beta^k(0w)$ have the same cardinality. Indeed, $\mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_1}) \setminus 0^m\Sigma_\beta^* = \mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_1}) \setminus 0^m\mathcal{L}_{m(\ell-1)}(\mathcal{A}_{\beta,q_1})$ and therefore $\#\mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_1}) \setminus 0^m\Sigma_\beta^* = V_{m\ell,1} - V_{m(\ell-1),1} = U_{m\ell} - U_{m(\ell-1)}$.

It suffices to show that the map $w \mapsto \mathrm{val}_\beta^k(0w)$ is injective. We assume by contradiction the existence of $w, w' \in \mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_1}) \setminus 0^m\Sigma_\beta^*$ such that $\mathrm{val}_\beta^k(0w) = \mathrm{val}_\beta^k(0w')$. Consequently, $\mathrm{val}_\beta(w) = \mathrm{val}_\beta(w')$. To conclude we proceed as in the proof of Lemma 6.4 for $k = 1$. □

Proof of Proposition 5.2. We denote $\mathcal{L} = \mathcal{L}(\mathcal{A}_\beta^k) \cap \Sigma_\beta(\Sigma_\beta^m)^* \setminus (00^m\Sigma_\beta^* \cup 1p_k\Sigma_\beta^*)$.

(Existence): If $n \geq 0$ then there exists a unique integer $\ell \in \mathbb{N}$ such that $U_{m(\ell-1)} \leq n < U_{m\ell}$. From Lemma 6.5 we obtain a word $w \in \mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_1}) \setminus 0^m\Sigma_\beta^*$ such that $\mathrm{val}_\beta^k(0w) = n$ and we set $u = 0w \in \mathcal{L}$. If $n < 0$ then there exists a unique integer $\ell \in \mathbb{N}$ such that $-V_{m\ell,k} \leq n < -V_{m(\ell-1),k}$. From Lemma 6.4 we obtain a word $w \in \mathcal{L}_{m\ell}(\mathcal{A}_{\beta,q_k}) \setminus p_k\Sigma_\beta^*$ such that $\mathrm{val}_\beta^k(1w) = n$ and we set $u = 1w \in \mathcal{L}$.

(Unicity): Let us assume by contradiction the existence of $u, u' \in \mathcal{L}$ such that $\mathrm{val}_\beta^k(u) = \mathrm{val}_\beta^k(u')$. If $u \in 0\Sigma_\beta^*$ then by Lemma 6.5, $\mathrm{val}_\beta^k(u) \geq 0$ and therefore $\mathrm{val}_\beta^k(u') \geq 0$. This implies $u' \in 0\Sigma_\beta^*$ by Lemma 6.4. Then by Lemma 6.5, $u = u'$. If $u \in 1\Sigma_\beta^*$ then by Lemma 6.4, $\mathrm{val}_\beta^k(u) < 0$. Therefore $\mathrm{val}_\beta^k(u') < 0$ and by Lemma 6.5, $u' \in 1\Sigma_\beta^*$. Consequently, by Lemma 6.4, $u = u'$. □

References

[1] J. Berstel. Fibonacci words - a survey. The book of L, dedic. A. Lindenmayer Occas. 60th Birthday, 13-27 (1986)., (1986).

[2] V. Berthé and M. Rigo, (eds.). *Combinatorics, automata and number theory*, volume 135 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, (2010).

[3] L. Carlitz. *Fibonacci representations.* Fibonacci Quart. **6** (1968), 193–220.

[4] D. E. Daykin. *Representation of natural numbers as sums of generalised Fibonacci numbers.* J. London Math. Soc. **35** (1960), 143–160.

[5] C. Frougny. *Linear numeration systems of order two.* Inform. and Comput. **77** (1988), 233–259.

[6] C. Frougny. *Fibonacci representations and finite automata.* IEEE Trans. Inform. Theory **37** (1991), 393–399.

[7] C. Frougny. *On-line finite automata for addition in some numeration systems.* Theor. Inform. Appl. **33** (1999), 79–101.

[8] C. Frougny and J. Sakarovitch. *Number representation and finite automata.* In 'Combinatorics, automata and number theory', volume 135 of *Encyclopedia Math. Appl.*, Cambridge Univ. Press, Cambridge (2010), 34–107.

[9] C. Frougny and B. Solomyak. *On representation of integers in linear numeration systems.* In 'Ergodic theory of $\mathbf{Z}^d$ actions (Warwick, 1993–1994)', volume 228 of *London Math. Soc. Lecture Note Ser.*, Cambridge Univ. Press, Cambridge (1996), 345–368.

[10] D. E. Knuth. *The art of computer programming. Vol. 1.* Addison-Wesley, Reading, MA, (1997). Fundamental algorithms, Third edition.

[11] D. E. Knuth. *The art of computer programming. Vol. 2.* Addison-Wesley, Reading, MA, (1998). Seminumerical algorithms, Third edition.

[12] S. Labbé and J. Lepšová. *A numeration system for Fibonacci-like Wang shifts.* In 'Combinatorics on words', volume 12847 of *Lecture Notes in Comput. Sci.*, Springer, Cham (2021), 104–116.

[13] C. G. Lekkerkerker. *Representation of natural numbers as a sum of Fibonacci numbers.* Simon Stevin **29** (1952), 190–195.

[14] M. Lothaire. *Algebraic Combinatorics on Words*, volume 90 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, Cambridge, (2002).

[15] A. Massuir, J. Peltomäki, and M. Rigo. *Automatic sequences based on Parry or Bertrand numeration systems.* Adv. in Appl. Math. **108** (2019), 11–30.

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

[17] J. Sakarovitch. *Easy multiplications. I. The realm of Kleene's theorem.* Inform. and Comput. **74** (1987), 173–197.

[18] J. Sakarovitch. *Elements of automata theory.* Cambridge University Press, Cambridge, (2009). Translated from the 2003 French original by Reuben Thomas.

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

[20] E. Zeckendorf. *Représentation des nombres naturels par une somme de nombres de Fibonacci ou de nombres de Lucas.* Bull. Soc. Roy. Sci. Liège **41** (1972), 179–182.

Evolutionary Equilibria in the Oligopolistic Market Respecting the Cost of Change

Veronika Lipovská

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

advisors:
Michal Červinka, Institute of Economic Studies
Faculty of Social Sciences, Charles University in Prague
Jiří Outrata, Department of Decision-Making Theory
Institute of Information Theory and Automation, CAS

Abstract. The contribution deals with the model of an oligopolistic market evolving over several time stages, where the companies adapt to changing input parameters while considering the cost of change of production. The Cournot-Nash equilibrium can be established and the algorithm for computing the corresponding vector of production strategies for each company is described.

Keywords: Cournot-Nash equilibrium, evolution, cost of change

Abstrakt. Tento příspěvek se zabývá modelem oligopolistického trhu, který se vyvíjí v průběhu několika období, v nichž se společnosti přizpůsobují proměnlivým vstupním parametrům a přitom berou v úvahu náklady na změnu produkce. Zde můžeme určit Cournotovo-Nashovo ekvilibrium a popsat algoritmus pro výpočet odpovídajícího vektoru produkcí pro každou společnost.

Klíčová slova: Cournotovo-Nashovo equilibrium, evoluce, náklady na změnu produkce

1 Introduction

The aim of this contribution is to describe a certain evolving oligopolistic market viewed by the optics of game theory and variational analysis. The game theory considers the encounters of agents, called *players*, and the sets of their possible behaviour, each instance called a *strategy*, leading to possible outcomes. As a result of an outcome, each player receives a (possibly negative) price. The aim of the players is to choose a strategy which minimizes their loss functions.

The theory is due to John von Neumann, who proposed the general framework between 1928 and 1941, leading to the joint work with Oskar Morgenstern in the book *Theory of Games and Economic Behaviour* [7].

In 1950, John Nash developed the concept of what is now known as *Nash equilibrium*, a state in which it does not pay off to unilaterally change the strategy.

The notion of equilibrium is much older, however. In 1838 Auguste Cournot used it for the market with two competing producers. The said market is called *duopoly*, later generalized for more players into *oligopoly.* The Nash equilibrium in such a market is known as *Cournot-Nash equilibrium.*

One of the many models the Cournot-Nash equilibria can be sought for is due to Sjur Didrik Flåm, [5], who proposes to repeat the game in discrete time steps and consider the cost related to each change of production, leading to a finite sequence of equilibria. Jiří Outrata suggests considering this game not in each time step separately, but as an *evolutionary equilibria* over the planning horizon of several time stages. This contribution describes the case when the players know what the parameters will be and plan accordingly.

The paper is structured as follows. Section 2 describes the model of the market, followed by Section 3 with the necessary background from the modern variational analysis. Section 4 briefly explains the splitting methods and the forward-backward splitting method that is used in the computation. Section 5 describes the computatition itself and contains the algorithm. Finally, Section 6 gives the parameters and concrete functions to the model. These were taken from [9], where the equilibria are computed separately for each time step.

Throughout the text, the following notation is employed. $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is a set-valued mapping, $\operatorname{dom} F$ is its domain, $\overline{\mathbb{R}} = \mathbb{R} \cup \infty$ is the extended real line, for a cone K, K° signifies its (negative) polar cone and $\underset{A}{\to}$ denotes the convergence within a set A.

2 Model

In this section, we properly define the above-mentioned notion of Cournot-Nash equilibrium and apply it to a specific model of oligopolistic market respecting the cost of change.

Definition 2.1. A (non-cooperative) *Cournot-Nash equilibrium* of an oligopolistic market with l players is a vector

$$(\bar{x}_1, \ldots, \bar{x}_l) \in A_1 \times \cdots \times A_l \text{ such that } \bar{x}_i \in \operatorname*{argmin}_{x_i \in A_i} g_i(x_i, \bar{x}_{-i}) \text{ for all } i,$$

where x_i is the production level of the ith player from his set of feasible strategies A_i, and $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_l)$ those of his rivals; $g_i(x_1, \ldots, x_l)$ is the objective function of the ith player.

The ith player then aims to minimize the function of s variables

$$\begin{aligned} &\text{minimize} \sum_{t=1}^{s} \left(c_i^t(p_1^t, x_i^t) - \langle x_i^t,\, \pi^t(p_2^t, T^t) \rangle + \beta_i^t \| x_i^t - x_i^{t-1} \| \right) \\ &\text{subject to} \\ &\qquad x_i^t, \in A_i^t,\ t = 1, \ldots, s, \end{aligned} \tag{1}$$

where x_i^0 is the initial production of the ith player and each production level x_i^t is chosen from the set $A_i^t \subset \mathbb{R}^n$ of feasible strategies of the player, n is the number of commodities in the portfolio. Further, $c_i^t : \mathbb{R}^{m_1} \times \mathbb{R}^n \to \mathbb{R}$ represent production costs, $T^t = \sum_{i=1}^l x_i^t$ is the total production vector of commodities, the inverse demand function $\pi^t : \mathbb{R}^{m_2} \times \mathbb{R}^n \to \mathbb{R}^n$ assigns to each value of parameter p_2^t and the vector T^t the vector of prices at which the consumers are willing to demand. In addition, $\beta_i^t \|x_i^t - x_i^{t-1}\|$ is the cost of change from production x_i^{t-1} to x_i^t, where $\|\cdot\|$ is an arbitrary norm and β_i^t are non-negative constants, all at time t.

Furthermore, we impose the following assumptions:

(S1) There exist open sets $\mathcal{B}_1^t \subset \mathbb{R}^{m_1}$ and open sets $\mathcal{D}_i^t \supset A_i^t$ for all $i = 1, \dots, l$, $t = 1, \dots, s$, such that

- c_i^t are twice continuously differentiable on $\mathcal{B}_1^t \times \mathcal{D}_i^t$;
- $c_i^t(p_1^t, \cdot)$ are convex for all $p_1^t \in \mathcal{B}_1^t$.

(S2) There exist open sets $\mathcal{B}_2^t \subset \mathbb{R}^{m_2}$ for all $t = 1, \dots, s$, such that

- π^t is twice continously differentiable on $\mathcal{B}_2^t \times \operatorname{int} \mathbb{R}_+$ and $\pi^t(p_2^t, \cdot)$ is strictly convex on $\operatorname{int} \mathbb{R}_+$ for all $p_2^t \in \mathcal{B}_2^t$;
- $\vartheta \pi^t(p_2, \vartheta)$ is a concave function of ϑ for all $p_2 \in \mathcal{B}_2$.

(S3) Sets A_i^t, $i = 1, \dots, l$, $t = 1, \dots, s$ are closed bounded intervals and at least one of them belongs to $\operatorname{int} \mathbb{R}_+$.

Under these assumptions, it is the last term of the function that is responsible for the nonsmoothness of the objective function. Fortunately, this nonsmooth term does not depend on the production of the other players, only on the previous productions of the player in question.

The assumptions further ensure that the famous Nash theorem for the existence of an equilibrium can be applied:

Theorem 2.1 (Nash)**.** Suppose that the sets of feasible strategies A_i are convex and compact and that for each $i = 1, \dots, l$ the objective functions g_i are continuous and the functions $x_i \to g_i(x_i, x_{-i})$ are convex. Then there exists a Nash equilibrium.

By means of variational analysis, the uniqueness of the solution to the problem (1) can be shown and the said solution computed.

3 Background from variational analysis

For readers' convenience, we review some basic notions of modern variational analysis.

Definition 3.1. Let A be a closed set in $\mathbb{R}^n$ and $\bar{x} \in A$. Then

$$T_A(\bar{x}) = \operatorname*{Limsup}_{t \searrow 0} \frac{A - \bar{x}}{t} = \{d \in \mathbb{R}^n \mid \exists\, d_k \to d,\, t_k \searrow 0 : \bar{x} + d_k t_k \in A \;\forall\, k \in \mathbb{N}\}$$

is the *tangent (contingent, Bouligand) cone* to A at $\bar{x}$,

$$\hat{N}_A(\bar{x}) = (T_A(\bar{x}))^\circ$$

is the *regular (Fréchet) normal cone* to A at $\bar{x}$, and

$$N_A(\bar{x}) = \operatorname*{Limsup}_{x \xrightarrow[A]{} \bar{x}} \hat{N}_A(x) = \{v \in \mathbb{R}^n \mid \exists\, x_k \xrightarrow[A]{} \bar{x},\ v_k \in \hat{N}_A(x_k) \text{ such that } v_k \to v\}$$

is the *limiting (Mordukhovich) normal cone* to A at $\bar{x}$,
where the 'Limsup' stands for the *outer set limit* in a sense of Painlevé and Kuratowski.

If the set A is convex, the normal cones coincide.

To state the optimality conditions, we need the subdifferential. The notion of subdifferentials is much broader, though for a convex function it can be simply stated as follows.

Definition 3.2. Consider a (single-valued) convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $\bar{x}$ with $f(\bar{x})$ finite. Then

$$\partial f(\bar{x}) = \{v \mid f(x) \geq f(\bar{x}) + \langle v, x - \bar{x}\rangle, \text{ for all } x \in \mathbb{R}^n\}$$

is the *(convex) subdifferential* of a function f at $\bar{x}$.

The problem of minimizing the convex function g over the set A is equivalent to solving the generalized equation (GE)

$$0 \in \partial g(x) + N_A(x).$$

For a monotone mapping T, its resolvent $J_{cT} = (I + cT)^{-1}$ with constant c is a single-valued nonexpansive function. If T is, in addition, maximal monotone, the domain of its resolvent is $\mathbb{R}^n$.

Further, for every positive λ the set of fixed points of the resolvent is equal to the set of zeros of the mapping, i.e. $x = J_{\lambda T}(x)$ if and only if $0 \in T(x)$.

4 Splitting methods

Splitting methods are used for seeking solutions to a generalized equation $0 \in T(x)$, where $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a monotone mapping which can be written as the sum of two other mappings $T = A + B$. We apply these methods in the situation when it is possible to find the resolvent of at least one of the mappings A and B, even though the resolvent of T might be difficult to obtain.

The forward-backward (FB) splitting method can be used for the case when A is maximal monotone and B single-valued. The solution to the GE $0 \in T(x)$ can be found via the following iteration: with $x^k \in \operatorname{dom} A$,

$$x^{k+1} = J_{c_k A}((I - c_k B)(x^k)). \tag{2}$$

It is easy to see that $0 \in T(x)$ is equivalent to $-cB(x) \in cA(x)$ and again to $x = J_{cA}((I - cB)(x))$. Thus the formula (2) is just a fixed point iteration with a varying multiplier c_k at each step.

The method got its name because in each iteration it performs the 'forward step', computing the value $(I - c_k B)(x^k)$ and the 'backward step', computing the resolvent of A.

5 Computation of optimal strategies of the players

The problem of minimizing the cost function of each player is equivalent to finding the zero of a generalized equation representing the subdifferential of the cost function. As the sum of smooth and nonsmooth terms, the GE can be written as the sum of the single-valued and the set-valued part

$$0 \in F_i(p,x) + Q_i(x_i),$$

where the set-valued term for the ith player amounts to

$$Q_i(x_i^1,\dots,x_i^s) = \sum_{t=1}^{s} \left(\Lambda_i^t(x_i^t - x_i^{t-1}) + N_{A_i^t}(x_i^t) \right),$$

with

$$\Lambda_i^t(\xi - a) = \begin{cases} \beta_i^t, & \xi > a, \\ -\beta_i^t, & \xi < a, \\ [-\beta_i^t, \beta_i^t], & \xi = a. \end{cases}$$

Furthermore, it can be shown that the terms Q_i are maximal monotone, therefore the FB splitting method is applicable to the problem.

It can be seen that, given some $c > 0$, the resolvent J_{cQ_i} of Q_i at an argument $(z^1,\dots,z^s)$ has the value

$$J_{cQ_i}(z^1,\dots,z^s) = (y^1,\dots,y^s),$$

where the vector $(y^1,\dots,y^s)$ is the unique solution to the optimization problem

$$\begin{aligned} &\text{minimize} \sum_{t=1}^{s} \left(\frac{1}{2}(y^t)^2 - z^t y^t + c\beta_i^t \|y^t - y^{t-1}\| \right) \\ &\text{subject to} \\ &\qquad y^t, \in A_i^t,\ t = 1,\dots,s, \end{aligned} \tag{3}$$

The algorithm then can be stated as follows:

```
initialization: $\varepsilon > 0$, $c > 0$, $k = 0, x^0 = (x_1^0, x_2^0, \dots x_l^0) \in (A_1) \times (A_2) \times \cdots \times (A_l)$,
if dist$(-F_i((x_i^1,\dots x_i^s)^k),\ Q_i((x_i^1,\dots,x_i^s)^k)) \le \varepsilon$ for all $i = 1,2,\dots,l$ then
    stop
end if
compute $z^k = x^k - cF_i(x^k)$    (forward)
for $i = 1,2,\dots,l$ do
    solve problem (3), arriving at the solution $y_i = (y_i^1,\dots,y_i^s)$    (backward)
end for
set $x^{k+1} = y$, $k = k+1$ and go to 2.
```

For the solution to the problem (3) the implementation uses the built-in Matlab function `fmincon`.

6 Example

Let us now consider an example from [9], the market with one product (n=1), five companies (l=5) and three periods (s=3). The feasible sets are the following intervals constant in time $A_1^t = [1, 150]$, $A_2^t = \cdots = A_5^t = [0, 150]$, $t = 1, 2, 3$, and the cost functions are of the form

$$c_i^t(b_i^t, x_i^t) = b_i^t x_i^t + \frac{\delta_i}{\delta_i + 1} K_i^{-\frac{1}{\delta_i}} (x_i^t)^{\frac{1+\delta_i}{\delta_i}},$$

where the parameters δ_i and K_i are constant in time and taken from [8, Table 12.1], listed in Table 1, and only parameters b_i^t change in time t. The values of b_i^t are listed in Table 2.

Table 1: Values of constant production parameters for companies.

	Firm 1	Firm 2	Firm 3	Firm 4	Firm 5
K_i	5	5	5	5	5
δ_i	1.2	1.1	1.0	0.9	0.8

Table 2: Values of parameters b_i^t changing in time.

b_i^t	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
$t = 1$	9	7	3	4	2
$t = 2$	10	8	5	4	2
$t = 3$	11	9	8	4	2

The cost of change $\beta_i^t \|x_i^t - x_i^{t-1}\|$ will appear only at production of companies 1, 2 and 3 with different time invariant constants

$$\beta_1 = 0.5, \quad \beta_2 = 1, \quad \beta_3 = 2.$$

Further, the market is characterized by the inverse demand function

$$\pi^t(\gamma, T^t) = 5000^{\frac{1}{\gamma}} (T^t)^{-\frac{1}{\gamma}}, \tag{4}$$

where γ is a positive parameter termed *demand elasticity* and $T^t = \sum_{i=1}^5 x_i^t$ is the total supply of the commodity at time t.

The numerical computations in Matlab are in progress.

7 Conclusion

In this contribution, we study a model of a certain oligopolistic market during several stages, leading to evolutionary equilibria, the existence of which is given by Nash theorem, the uniqueness can be proved by methods of second-order calculus of variational analysis. This is one of the results in the article in preparation [3].

This is work in progress and when the computations are finished, the results will be compared to those of [9], where the authors compute equilibria for each stage separately for the same market and data.

The FB splitting method used in this paper, though suitable for low dimensions such as in our example, will have to be replaced with other methods in case of more complex models.

References

[1] J.-P. Aubin. *Optima and Equilibria, An Introduction to Nonlinear Analysis*. Springer, corr. 2nd ed., (1998). ISBN: 978-3540649830.

[2] A. A. Cournot. *Researches into the Mathematical Principles of the Theory of Wealth.* The Macmillan Company, New York, (1838, English translation 1897).

[3] M. Červinka, V. Lipovská, J. Outrata. *Multistage Strategies for Players of Oligopolistic Markets Respecting the Cost of Change.* (2022), article in preparation.

[4] F. Facchinei, J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume II.* Springer Science+Business Media, New York, (2003). ISBN: 978-0-387-95581-0.

[5] S. D. Flåm. *Games and Cost of Change.* Annals of Operations Research **301** (2021), 107–119.

[6] J. Nash. *Non-cooperative Games.* Annals of Mathematics **54** (1951), 286–295.

[7] J. von Neumann, O. Morgenstern. *Theory of Games and Economic Behaviour.* Princeton University Press, Princeton, (1944).

[8] J. Outrata, M. Kočvara, J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*. Kluwer Academic Publishers, Boston, (1998). ISBN: 978-1-4757-2825-5.

[9] J. V. Outrata and J. Valdman. *On Computation of Optimal Strategies in Oligopolistic Markets Respecting the Cost of Change.* Math. Meth. Oper. Res. **92** (2020), 489–509.

Analogue of Mandelbrot set in Quantum Purification Protocols with Higher Degree[*]

Martin Malachov

7th year of PGS, email: malacmar@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. This paper presents latest result about fractal structures formed by quantum purification protocols of higher order evolution functions. Such protocols are important for maintaining quantum information and communication practically realisable. Several interesting conjectures are proposed based on calculations of the fractal dimensions.

Keywords: fractal, dimension, quantum information and communication

Abstrakt. Tento článek prezentuje poslední výsledky o fraktálních strukturách generovaných purifikačními protkoly s evolučními funkcemi vyšších řádů. Takové protkoly jsou nezbytné pro udržení kvantové informace a komunikace v spolehlivém režimu praktického užití. Na základě numerických výpočtů fraktálních dimenzí formulujeme několik zajímavých hypotéz.

Klíčová slova: fraktál, dimenze, kvantová informace a komunikace

1 Introduction

Quantum information and communication represent modern technological branches of science with a promise to change the world. Still, much remains to be done before quantum computers would a become common part of our lives. One of the biggest problem of quantum states to be successfully kept and manipulated is the decoherence. The inescapable interaction with the environment and the quantum nature of the physics cause the pieces of information, qubits, to be damaged. Few algorithms have been proposed, e.g. analogues to error correcting codes to keep the quantum information and communication realisable.The algorithms may rely on multiple copies sent with a control mechanism that can detect the state disruption.

Our focus will lie in a different type of algorithm which was in its fundamental form proposed in [1], formally transferred to arbitrary dimensions in [2] and further investigated in [3]. In the last paper it was shown that the nonlinear character of the protocol application causes emergence of the chaos in its pure form, easily understood via sensitive dependence to initial conditions. One of the typical features of the chaos is revelation of fractal shaped structures. We have previously [4] proposed single qubit version of the protocol and have studied the fractal structure properties when applied to general

[*] This work has been supported by project CAAS with registration number CZ.02.1.01/0.0/0.0/16_019/0000778.

mixed states. Later, we have introduced [5] modifications to the original protocol by so called twirling gates. This enhancement gave rise to wide new sets of fractals showing the chaos evolution can have many of theoretically possible regimes in an actual physical implementation. Now, we show another generalisation of the protocol that naturally follows previous studies. After defining the protocol and its action we briefly review basic knowledge, then we propose new generalisations and focus on the features that remain same as well as those that do differ from previously known state of art. In the end, we briefly introduce concept of Mandelbrot set and generalise it to our case; the main result being the pictures of Mandelbrot sets and their comparisons based on the parameters of our generalised protocols.

2 Chaotic protocol

We would like to introduce the chaotic protocol now but we essentially need to formalise the concept of qubit first. It can be realised by any two-level physical system, e.g. photons with two possible states of polarisation (horizontal, vertical). State of such a system can be in its most general conditions described via density operator; given the so called computational basis marking qubit states $|0\rangle, |1\rangle$ we can express the operator with the matrix in the computational basis:

$$\rho = \frac{1}{2}\begin{pmatrix} 1+w & u+iv \\ u-iv & 1-w \end{pmatrix}; u, v, w \in \mathbb{R}, u^2+v^2+w^2 \leq 1 \tag{1}$$

This form suggests geometrical interpretation of the qubit as a point in a three-dimensional ball. The protocol of our consideration manifests as formula containing elementwise product=:

$$\rho \to \rho' \equiv \frac{U(\rho \odot \cdots \odot \rho)U^\dagger}{\mathrm{Tr}(U(\rho \odot \cdots \odot \rho)U^\dagger)} \tag{2}$$

The matrix U is called twirling gate and it allows us to modificate the protocol action. The original protocol used the Hadamard gate $U = H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ but we consider all possible twirling gates and suggest to parameterise them in following way: $U = TR$ where

$$T_\tau = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\tau} \end{pmatrix}, \quad R_{x,\psi} = \begin{pmatrix} \cos x & \sin x e^{i\psi} \\ -\sin x e^{-i\psi} & \cos x \end{pmatrix} \tag{3}$$

with angles $x, \psi, \tau \in [0, 2\pi)$. Of course, in the most general case such matrix can be multiplied by any complex unit $e^{i\omega}$ but such global phase has no meaning for the physical state, therefore we omit it. Earlier, in [5], we have derived evolution equations in terms of w, u, v and we have shown that the asymptotic dynamics is equivalent for two protocols with x_1, ψ_1, τ_1 and x_2, ψ_2, τ_2 when they satisfy:

$$x_1 = x_2 \wedge \psi_1 - 2\tau_1 = \psi_2 - 2\tau_2 \tag{4}$$

The meaning of the equivalence lies in the fractal structure which is the same for such operators and we gave unique relationship among the attractors of the equivalent protocols. This fact reduced the set of all possible twirling operators (generally depending on three angles) to only two effective angles. Other symmetries have also been detected and discussed in [5].

Now, we introduce further generalisation of the protocol. The original protocol used one copy of the qubit to modify another copy. We suggest to use more copies, the scheme of the new protocol is in figure 1. The key element is the cluster of CNOT operator which are quantum version of the classical CNOT gates; such gate flips (or not) the value of the target bit when the control bit is in state 1 (or 0). In quantum version, the states of computational basis $|0\rangle, |1\rangle$ are flipped in the qubit expressed as a superposition of the basis states. More about the quantum gates and their action can be found in [6] or other books dedicated to quantum information. The projections noted in the scheme perform the flip and in this way they represent measuremenet-based selection. The projection can be generally performed onto both basal states $|0\rangle, |1\rangle$ but the projection on the latter states induces dull evolution where all states are mapped to a single state. Such degenerated protocol has no relevant practical use (except for some possible reset of the quantum state) and so we use projections of all states onto $|0\rangle$ in the rest of this paper. The evolution equations of the general qubit follow but first we introduce factors

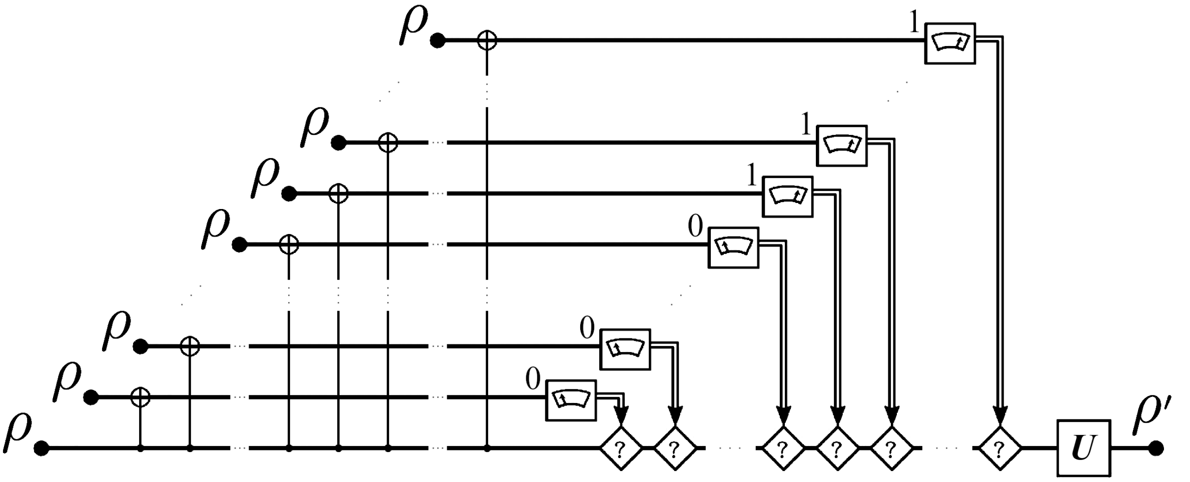


Figure 1: Scheme of the generalised purification protocol. $n-1$ copies of qubit ρ are used to repair the n-th copy via CNOT operators followed by measurement. The nonlinear (even chaotic) behaviour arises from this measurement-based selection.

$$N_n = \sum_{k \text{ even from } 0}^{n} \binom{n}{k} w^k \quad , \quad W_n = \sum_{k \text{ odd from } 1}^{n} \binom{n}{k} w^k \quad , \tag{5}$$

$$U_n = \mathscr{Re}\,(u+iv)^n \quad , \quad V_n = \mathscr{Im}\,(u+iv)^n \tag{6}$$

that describe the nonlinear mapping imposed by the CNOT gates. The twirling operators responsible for the geometrical modification remain in the same way as in the original protocol and for that reason the evolution equations are

$$
\begin{aligned}
u' &= \begin{cases} -\dfrac{W_n}{N_n}\sin 2x\cos(\psi-\tau)+ \\ \qquad +\dfrac{U_n}{N_n}\left[\sin\psi\sin(\psi-\tau)+\cos 2x\cos\psi\cos(\psi-\tau)\right]+ \\ \qquad +\dfrac{V_n}{N_n}\left[-\cos\psi\sin(\psi-\tau)+\cos 2x\sin\psi\cos(\psi-\tau)\right] \end{cases} \\
v' &= \begin{cases} -\dfrac{W_n}{N_n}\sin 2x\sin(\psi-\tau)+ \\ \qquad +\dfrac{U_n}{N_n}\left[-\sin\psi\cos(\psi-\tau)+\cos 2x\cos\psi\sin(\psi-\tau)\right]+ \\ \qquad +\dfrac{V_n}{N_n}\left[\cos\psi\cos(\psi-\tau)+\cos 2x\sin\psi\sin(\psi-\tau)\right] \end{cases} \\
w' &= \begin{cases} \dfrac{W_n}{N_n}\cos 2x+ \\ \qquad +\dfrac{U_n}{N_n}\cos\psi\sin 2x+ \\ \qquad +\dfrac{V_n}{N_n}\sin\psi\sin 2x \end{cases}
\end{aligned}
\tag{7}
$$

There is a key observation that the twirling gates responsible for the geometrical factors manifest the same way regardless of n. For this reason we can straightforwardly adopt the proposition of the asymptotic equivalence and omit the parameter τ that can be included into ψ via relation $\psi_2 = \psi_1 - 2(\tau_2 - \tau_1)$.

In this paper we restrict ourselves to study the evolution of the pure states so we can implement the theory of complex functions and discuss the analogy of the Mandelbrot set. For that reason we now remind that pure states form the Bloch sphere $u^2 + v^2 + w^2 = 1$. Such sphere can be identified with the Riemann sphere of complex numbers. One of the easy ways to do so lies in the geometrical intuition again. Performing stereographical projection with respect to the south pole of the Bloch sphere, i.e. state $|1\rangle$: $z = \frac{u+iv}{1+w}$. In this way the pure state qubit can be described as a two-dimensional complex vector $\begin{pmatrix} 1 \\ z \end{pmatrix}$, more precisely a ray of projective space CP^1, therefore depending only on one complex parameter z. The evolution of the pure state than manifests as evolution map

$$
z' = f_{n,x,\psi}(z) = \frac{z^n \cos x - e^{-i\psi}\sin x}{\cos x + z^n e^{i\psi}\sin x} \tag{8}
$$

which is rational polynomial function for all considered twirling operators and the number of the qubit copies used. The number of the copies determines the degree of the polynomials, the degree of the rational polynomial function.

We conclude we have families of protocols that depend in general on two parameters - angles - of the twirling gate and the number n marking the degree of the function and determined by the number of $n-1$ qubit copies used to modify the n-th one. All the protocols manifest as rational polynomial functions of a single complex variable. For such we can exploit the accessible theory sketched in next section.

3 Mandelbrot set

The theory of complex functions of a single complex variable grew to a wide field in the last century thanks to names of Fatou, Julia and Mandelbrot. For the spatial reasons we cannot give many of the fundamental theorems and refer to books like [7, 8].

The main conclusion drawn from the literature for our functions defined on the domain equal to the Riemann sphere follows: all the complex numbers are divided into two disjoint sets of, vaguely spoken, regular points called the Fatou set of the function and the points exhibiting chaotic behaviour called the Julia set of the function. The union of these two sets forms the whole Riemann sphere. The shape of the Julia set is often fractal but of course, it can be also a common set like circle. For the exact definitions and overview of the properties of the Julia and Fatou set, please refer to [7].

We determine the Julia set structure numerically as the analytical approach is impossible. After calculating the evolution of a grid of points we detect borders among regions of states with the same asymptotical regime, these regions are called basins of attraction. The states forming borders among the basins then belong to the s/et of our interest. Being often a fractal we calculate its dimension d using so called box-counting method [9]. For each family of the function of order n depending on parameters x, ψ we obtain a single number, dimension of the borders of the basins of attractions. In this way we obtained a map for each n: $(x, \psi) \to d$. The reason to exclude the parameter n lies in its physical and fundamentally important meaning. The order n (intuitively) has marcant influence on the speed of convergence. Yet it is unclear how it modulates the fractal shapes.

Now we introduce the Mandelbrot set. It can be defined in two equivalent forms. Consider all complex functions of form $f_c(z) = z^2 + c$ where c is a complex parameter. Each such function creates so called filled Julia set which consists of points not diverging to infinity which is critical attractive point for all such functions. It can be proven that the filled Julia set of such functions is either path-connected or totally disconnected with the Julia set being its border. We define Mandelbrot set

$$\mathcal{M}_f = \big\{c \in \mathbb{C} | f_c(z) = z^2 + c \text{ has connected Julia set}\big\} \tag{9}$$

The connected case automatically means that the Julia set has dimension equal at least to 1. Unfortunately, there can be totally disconnected sets constructed in a way to have arbitrary dimension. For example set of rational numbers is totally disconnected but has dimension 1 in the standard topology of $\mathbb{R}$. For this reason we cannot set up the Mandelbrot set equivalently with the definition based on dichotomy of the dimension of Julia set but generally for more general families of functions

$$\mathcal{M}'_f = \{c \in \mathbb{C} | f_c \text{ has Julia set with dimension } \geq 1\} \tag{10}$$

However, the set $\mathcal{M}'(f)$ possesses one crucial property: it can be numerically approximated by calculating dimensions of the corresponding fractals via box-counting method. We are also interested in the actual value of the dimension because it meaures (in its way) the complexity of the chaotic behaviour. Functions with the Julia set of dimension 2 are important for that reason that each point (state) belongs to the Julia set and thus is sensitive to the smallest perturbations.

Alternative definition of the Mandelbrot set is available thanks to Mandelbrot himself as $\mathcal{M}_f = \{c \in \mathbb{C} | \{f^{\circ n}(0) | n \in \mathbb{N}\} \text{ is bounded}\}\}$. Alternatively saying the iterations of critical point 0 do not converge to infinity. However, this approach is not viable for rational polynomial functions of our consideration as there is no such pair of points with the same properties guaranteed. Also the theorem linking the evolution of the critical points with the connectedness of the Julia set is no more valid. The formal approach to handle the Julia sets of rational polynomial functions lies in their definition as a closure of all repelling periodic points. Such definition does not admit any easy classification neither with respect to connectedness nor to orbits of critical points. For these facts we restrict to the dimension as a sufficeint, available and useful characteristic of the function.

To make tise context of the Mandelbrot set generalisation more reliable despite the mentioned problems mentioned, we remark that our evolution functions 8 depend on two real variables that can be composed into a single complex constant $C = e^{i\psi} \tan x \in \mathbb{C}$ though value of $x = \frac{pi}{2}$ must be discussed separately then. For this reason we deal with a single-parametric family of functions just like the Mandelbrot set.

4 Fractal dimension for generalised protocols

The analysis of pure states evolution clearly has following properties: the Julia set has dimension equal at most to 2. Such case means that each qubit belongs to the Julia set meaning it is sensitive to initial conditions. The protocol than loses its practical use for entanglement distillation, where entangled states can be repaired after its perturbance. However, one can efficiently use the protocol to create a random qubit. The evolution of a generic state belonging to the Julia set runs densely through the Julia set - all states. Of course, many iterations should be used to allow two initially closed states to be separated.

The calculation of the fractal dimension is performed by a selfmade script in Matlab environment. The box-counting dimension [9] is computer-friendly simplified version of the original Hausdorff dimension expressed in terms of optimal coverings. The covering is instead performed by a grid of frames (pixels or group of pixels) to count the sets covering the object.

First we make few notes on the case of $n = 2$, where the original protocol [1] takes place with $x = \frac{\pi}{4}, \psi = 0$. The fractal induced by corresponding $f_{x,\psi}$ has dimension $\doteq 1.55$ and this dimension changes only slightly and continuously in the parameter neighbourhood.The modulation of f also changes the attractor cycles yet for practical (i.e. experimental) application the sufficiently small perturbance to the protocol execution does not yield drastic change of the result as could be expected from the chaotic nature of the protocols. The exact description of this type of stability is not in the scope of this work and we move on to the topic of Mandelbrot set.

The reason to take only parameters $\in [0, \frac{\pi}{2}]$ is in the inner symmetries, e.g. x and $x + \pi$ give the same function which can be seen easily from 7.

The result of calculated relationship $x, \psi \to d$ for $n = 2$ is shown in figure 2. While there seem to be regions where the dimension changes only slightly and smoothly, like large regions of x close to 0 or $\frac{\pi}{2}$ where the fractal is formed by a closed curve with $d \doteq 1$, region near $x = \frac{pi}{4}, \psi \doteq 0$ where the dimension evaluates to $d \doteq 1.55$, or large region at $x = \frac{\pi}{4}, \psi \doteq \frac{\pi}{2}$ where the dimension reaches values close to 2. These regions

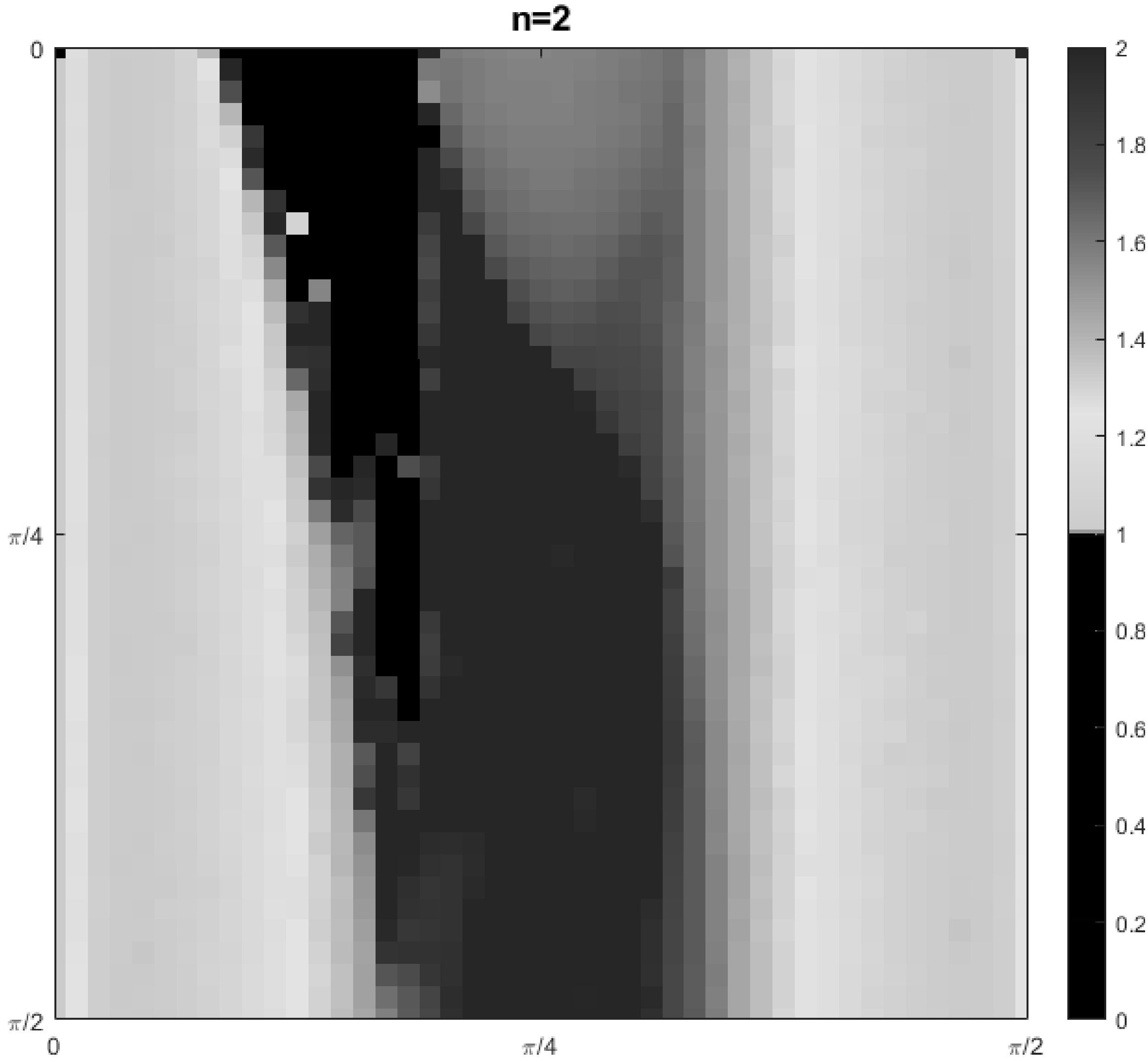


Figure 2: Overview of the dimension depending on the twirling angles x (horizontally), ψ (vertically from top to bottom). The dimension is represented by the colour coded in the scheme to the right. The generalised Mandelbrot set $M(f_{n,x,\psi})$ is the complement of the black region which marks parameters for which the box-counting method finds only a set of points (dimension $0 < d < 1$), typically signalling totally disconnected Julia set.

seem to be bordered by thin layers, possibly curves of sudden changes. We have to be aware of the fact that the box-counting is a numerical method burdened with error that can cast a bias to the results. Such bias is obvious for $x \doteq 0$ where the function is $f(z) \doteq z^2$ and the unit circle is distorted to a fuzzy shape. The fuzziness causes effect where neghbouring pixels are aslo counted to the covering while they would not be when working in higher resolution. The resolution and computational demands are the natural drawback of this method, therefore we didn't choose higher resolution of the image 2. Still, the benefits of the method are far more valuable at this moment where we check qualitative characteristics. And being aware of the complications we can avoid misinterpretation of the results. One of the misinterpretation would be that the black region of seemingly dimension zero means that the Julia set is empty. From the theory [7] the Julia set for any rational function is nonempty. However, the region capture the Julia sets that are totally disconnected. The set of points cannot be well captured by the boxes of the box-counting method yielding number < 1 that does not have to match the true dimension and typically is much lower, close to 0. Yet our purpose was to show the generalised Mandelbrot set and the black region is the complement of $M'(f_{2,x,\psi})$. The impairment of the box-counting method actually turns out useful in this case of Julia sets with $d < 1$.

For the higher orders n we present an overview in figure 3. There are three crucial findings. First, there appears to be an additional symmetry in terms of ψ depending on the value of n. E.g. the shape found for $n = 3$ seems only slightly modified and copied three times in the image for n. Generally, our conjecture suggests there are $n - 1$ copies of $n = 2$ similar shape to n odd case, $n - 2$ copies of $n = 3$ similar shape for protocols of order n odd. To prove this metasymmetry in the function of dimension $d_n = d_n(x, \psi)$ is probably formidable task because no theoretical prescripton for the dimension can be given and the nummerical methods can never give a result precise enough. At the moment we cannot give any advice how to penetrate into the hard task of this conjecture.

The second finding is that the cases of odd n yield qualitatively different generalised Mandelbrot sets than those of even n. The box-counting method has not detected Julia sets with $d < 1$ for odd n. Such result would imply the Mandelbrot set is equal to the whole parameter space, still there is the uncertainty of the nummerical error as totally disconnected sets of dimension close to one can be assigned dimension 1 by the box-counting just as it could happen for the set of rational numbers in the line of real numbers.

Third crucial finding is that with increasing n the values of x which yield fractal shaped Julia sets concentrate near value $x = \frac{\pi}{4}$. The numerical value of the dimension for $x = \frac{\pi}{4}, \psi = 0$ has been determined to be very near (within the precision of the box-counting) to the value 1.55 (obtained for $n = 2$) for other values of n up to 8 too. We propose a conjecture that the dimension of the Julia set of the corresponding functions $f_{n,\pi/4,0}(z) = \frac{z^n-1}{z^n+1}$ is the same regardless of n. Again, without any analytic clue to the dimension value this task is uncrackable.

Last point of our results is that for $n = 2, x = \frac{pi}{4}, \psi = \frac{\pi}{2}$ the corresponding function $f(z) = \frac{z^2+i}{1+iz^2}$ can be shown to have Julia set equal to the whole Riemann sphere $\mathbb{C}$. Each state undergoes deterministic chaos. Such feature might repeat for even values of n though not for the odd values. This conjecture can be possibly grasped using tools of geometrical transformation and multiple covering of a thorus that are used to argument the proof for the $n = 2$ case, see [10].

5 Conclusion

The fact that the dimension of the Julia sets reflects the even-odd order is very interesting and definitely deserves further, hopefully analytical work though the task seems impossible to be grasped at the moment. The symmetry in terms of ψ when the order n is changed is also very interesting and can be possibly used in experimental applications when some angles (e.g. in terms of optical components or ion manipulation) are difficult to implement. Possibly only one of our conjectures can be actually proved with contemporary state of art. Still, a better numerical support for the conjectures would mean an important step in the field. The resources to fulfill such numerical work increase drastically with the precision, though.

One of the most interesting results is the sudden change of the dimension of the functions' Julia sets with small changes to the function parameters. This metachaotic (possibly exponentially sensitive behaviour of functions $f_{n,x,\psi}$ that induce exponentially sensitive response of the input variable z) behaviour of the chaotic functions has been already noted for the mentioned functions $f(z) = z^2 + c$ where the concept of J-stability has been proposed. We find it interesting now to try to reproduce or generalise the knowledge of J-stability to the context of rational polynomial functions of our interest.

We also support the experimental execution of the protocol to verify the results. The technology level of today's state of art is probably sufficent to perform sufficent number of iterations of the protocol.

References

[1] H. Bechmann-Pasquinucci et al. *Nonlinear Quantum State Transformation of Spin 1/2.* Phys. Lett. A **242** (1998), 198–204

[2] G. Alber, A. Delgado, N. Gisin, I. Jex. *Efficient Bipartite State Purification in Arbitrary Dimensional Hilbert Spaces.* J. Phys. A: Math. Gen.**34** (2001), 8821

[3] T. Kiss, I. Jex, G. Alber, S. Vymětal. *Complex Chaos in the Conditional Dynamics of Qubits.* Phys. Rev. A **74** (2006), 040301(R)

[4] M. Malachov, I. Jex, O. Kálmán, T. Kiss. *Phase Transition in Iterated Quantum Protocols for Noisy Inputs.* Chaos **29** (2019), 033107. arXiv: 1809.00140

[5] M. Malachov *Manipulation of Time Evolution in Quantum Purification Protocol* Doktorandské dny 2019 conference proceedings (2019)

[6] A.M. Nielsen, I.L. Chuang. *Quantum Computation and Quantum Information.* Cambridge University Press (2000)

[7] J.W. Milnor. *Dynamics in One Complex Variable*, 3^{rd} edition. Princeton University Press (2000)

[8] K. Falconer. *Fractal Geometry: Mathematical Foundations and Applications*, 2^{nd} edition. Wiley (2003)

[9] S.H. Strogatz. *Nonlinear Dynamics and Chaos*, 2^{nd} edition. Westview Press (2015)

[10] A. Gilyén, T. Kiss, I. Jex. *Exponential Sensitivity and its Cost in Quantum Physics* Scientific Reports **6** (2016), 29976. arXiv: 1508.0319

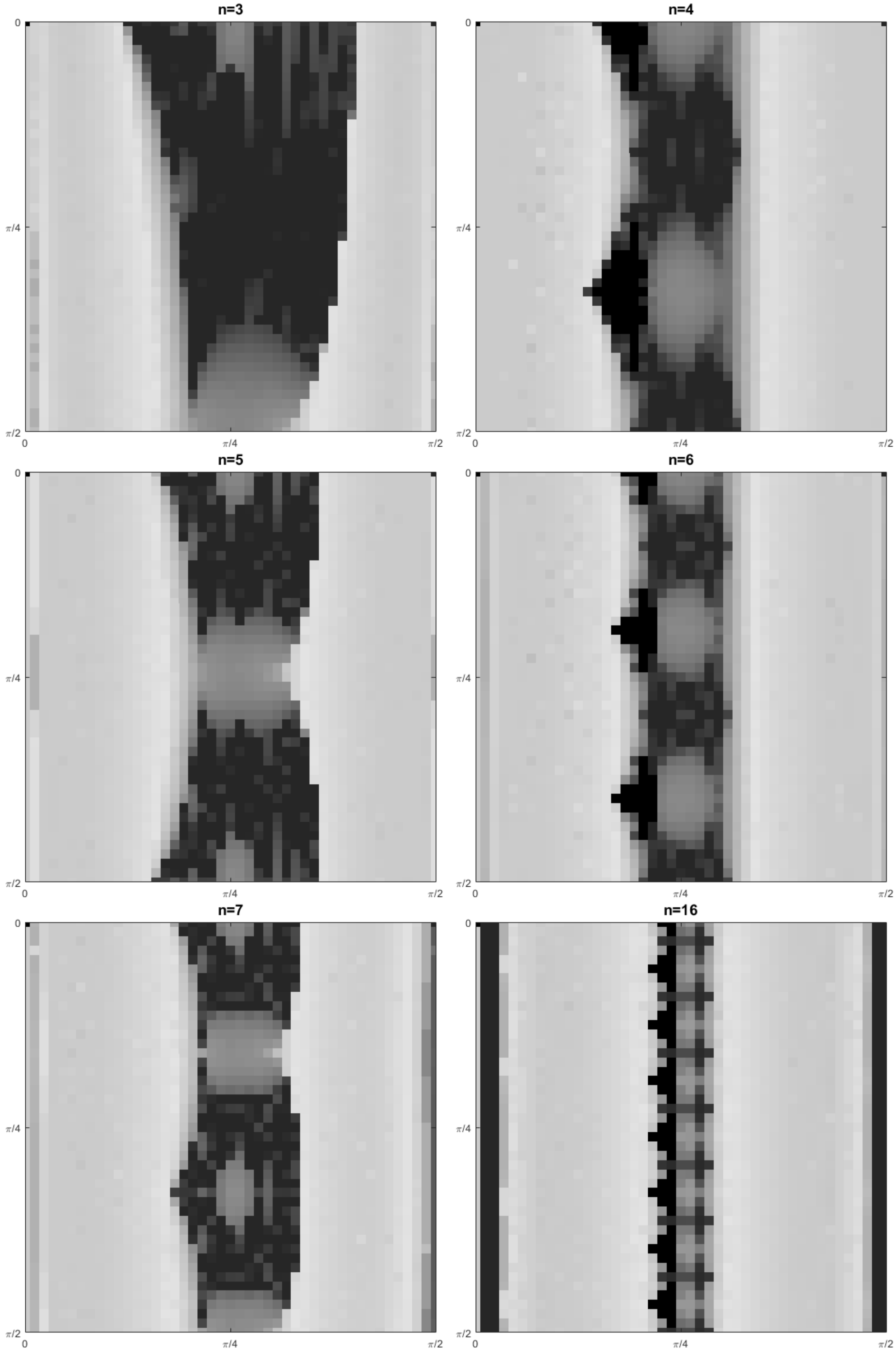


Figure 3: Overview of the generalised sets $M(f_{n,x,\psi})$ (in black) for higher orders n. Setting of the axes and the colormap are the same as in figure 2. Numerical bias near $x = 0, x = \frac{\pi}{2}$.

Nondegenerate Homotopy and Geometric Flows

Jiří Minarčík

5th year of PGS, email: jiri.minarcik@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 addresses the problems associated with formulations of geometric flows which depend on the existence of the Frenet frame. In order to better understand the limitations of such motion laws and to predict their long-term behavior, we introduce new quantity which is invariant to nondegenerate homotopies and use it to classify the space on which these motion laws operate.

Keywords: geometric flows, locally convex curves, Frenet frame, nondegenerate homotopy

Abstrakt. Tento příspěvek řeší problémy spojené s formulací geometrických toků křivek založenou na existenci Frenetova repéru. Abychom byli schopni lépe pochopit omezení těchto pohybových zákonů a uměli predikovat jejich dlouhodobý vývoj, zavedeme novou veličinu, která je invariantní vůči nedegenerované homotopii. Její hodnota nám poslouží ke klasifikaci prostorů, na nichž tyto pohybové zákony operují.

Klíčová slova: geometrický tok, lokálně konvexní křivky, Frenetův repér, nedegenerovaná homotopie

References

[1] J. Minarčík, M. Beneš. *Minimal surface generating flow for space curves of non-vanishing torsion.* Discrete & Continuous Dynamical Systems - B, 2022(27), 6605–6617.

[2] J. Minarčík, M. Beneš. *Nondegenerate homotopy and geometric flows.* Homology, Homotopy and Applications, 2022(24), 255–264.

Learning State Correspondence of Reinforcement Learning Tasks for Knowledge Transfer

Marko Ruman

5th 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. Successful and widely known works such as DQN or AlphaGo made deep reinforcement learning widely popular field. They achieved super-human performance in solving complex tasks some of which were thought impossible to solve due to e.g. number of possible states. One of the biggest downsides, however, is still their inability to generalize across related tasks, which for humans is very natural. The generalizing and reusing knowledge from the past is a crucial part of a generally intelligent agent.

This work formalizes the problem of knowledge transfer in reinforcement learning and offer a novel method for one-to-one task knowledge transfer. The method makes use of GAN model which is tailored here specifically for reinforcement learning tasks. The method assumes unpaired data records from *source* and *target* reinforcement learning tasks containing current state, action and future state. Thus, the method learns how to transfer knowledge in an unsupervised way.

The work offers couple of experiments with Atari game Pong, where it demonstrates the potential of the proposed method as well as the difficulties that arise even when used on simple environments.

Keywords: deep reinforcement learning, transfer learning, Markov decision process

Abstrakt. Úspěšné a obecně známé práce jako DQN nebo AlphaGo se zasloužili o to, že hluboké zpětnovazební učení je aktuálně velmi populární vědní disciplína. Zmíněné metody se naučili řešit komplexní úlohy lépe než lidé a dokázali se dokonce naučit řešit úlohy, kde se to považovalo za nemožné například kvůli obrovského počtu možných stavů. Jednou z hlavních nevýhod těchto metod ale pořád zůstava jejich neschopnost zobecňovat naučené znalosti na podobné úlohy. Zobecňování a znovupoužití již naučených znalostí je nezbytná součást obecně inteligentních agentů.

Tato práce formalizuje problém přenosů znalostí ve zpětnovazebném učení a přináší novou metodu na přenos mezi 2 úlohami. Metoda používá model GAN, který je v práci modifikovaný pro použití pro zpětnovazební úlohy. Metoda přepodkládá dvě nespárované množiny datových záznamů ze *zdrojové* a *cílové* úlohy obsahující aktuální stav, akci a budoucí stav. Z toho plyne, že metoda se učí jak přenést znalosti mezi úlohami "bez učitele"(unsupervised learning).

Práce obsahuje vícero experimentů s Atari hrou Pong, kde se demonstruje potenciál přinášené metody a také problémy se kterými je možno se setkat i v tomto relativně jednoduchém prostředí.

Klíčová slova: hluboké zpětnovazební učení, přenos znalostí, Markovské rozhodovací procesy

Full paper: Ruman, M., & Guy, T.V. (2022). Learning state correspondence of reinforcement learning tasks for knowledge transfer. Under review in International Journal of Machine Learning and Cybernetics.

Matrix Form of Multidimensional Continued Fractions*

Hanka Řada

5th year of PGS, email: hanka.rada@gmail.com
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. In this work we study the problem of periodicity of multidimensional continued fractions (MCFs) using the matrix properties of these algorithms.

Using this approach, we show that there exist a class of vectors which can not have a purely periodic expansion in any unimodular weakly convergent MCF algorithm. We also discuss some possible reasons why most of the well-known MCF algorithms seems to fail to have periodic expansion for every base of a cubic number field.

Keywords: multidimensional continued fractions, periodicity, matrices of linear transformation

Abstrakt. V tomto textu se věnujeme algoritmům vícerozměrných řetězových zlomků a zejména otázce periodicity těchto algoritmů. Algoritmy prezentujeme v jejich maticové podobě a ukazujeme, že matice repetendu rozvojů v těchto algoritmech jsou rovny maticím jistých lineárních transformací.

Dále ukazujeme, že existuje třída vektorů, jež nemohou mít čistě periodický rozvoj v žádném unimodulárním slabě konvergentním algoritmu vícerozměrných řetězových zlomků. Diskutujeme také možné příčiny toho, proč se zdá, že u většiny klasických algoritmů vícerozměrných řetězových zlomků existuje báze kubického tělesa, která nemá periodický rozvoj.

Klíčová slova: vícerozměrné řetězové zlomky, periodicita, matice lineárních transformací

1 Introduction

In 1839 [6] Hermite asked Jacobi if there is an algorithm that would detect the algebraic degree of any algebraic number (for definitions of these terms se Preliminaries). For the rational numbers (numbers of degree 1) is such an algorithm the decimal expansion of a number. In the case of quadratic numbers (the numbers of degree 2) is such a system also well-known. Namely, it is the regular continued fraction representation. However, we still do not have a satisfactory answer for numbers of degree three and higher.

In order to solve this question, there were introduced many multidimensional continued fraction (or MCF for short) algorithms. We will focus only on the vectorial algorithms. That are the algorithms that can be written as a matrix multiplication. For more information about the other type of MCF algorithms, the geometric algorithms, see [7].

*The work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS20/183/OHK4/3T/14.

The first and the most often studied MCF algorithm is the Jacobi-Perron algorithm, introduced by Jacobi (1868, [5]) and later generalised by Perron (1935, [9]). Other well-known algorithms are Poincaré algorithm (1884, [10]), Brun algorithm (1920, [3]), Selmer algorithm (1961, [13]) and Fully subtractive algorithm (1995, [11]). There exist also many modifications of this algorithms, for example the Algebraic Jacobi-Perron algorithm (AJPA) by Tamura and Yasutomi (2009 [14]). For a detail description of the well-know MCF algorithms and their properties see the books [2] and [12]. A good overview of these MCF algorithms is also in [8].

In this work we present another approach to this problem. We study the algorithms in their matrix form. In the main theorem (Theorem 9) we state, that every matrix of repetend of a MCF expansion is equal to a matrix of some linear transformation. Moreover, in the third section of this text, we present a theorem, which gives us instructions, how to find these matrices.

Using these two main results, we show, that there exists a class of vectors which can not have a purely periodic expansion in any unimodular weakly convergent MCF algorithm. Moreover, we show (using an example), the reason, why we believe that most of the well-known MCF algorithms seems to fail to answer the Hermite question.

2 Preliminaries

A number $\alpha \in \mathbb{C}$ is *algebraic* (over $\mathbb{Q}$) if it is a root of some polynomial $f \in \mathbb{Q}[\alpha]$. The set of algebraic numbers (over $\mathbb{Q}$) is denoted by $\mathbb{A}$. Let α and α' be roots of the same irreducible polynomial f. We say that α' is a *conjugate* of α.

The *degree* of α is the least number n such that α is a root of a polynomial of degree n. Algebraic numbers of degree two are called *quadratic* and algebraic numbers of degree three are called *cubic* (they are roots of *quadratic* respectively *cubic* polynomial with rational coefficients).

Let $\alpha \in \mathbb{A}$. The *number field* $\mathbb{Q}(\alpha)$ is defined by

$$\mathbb{Q}(\alpha) := \bigcap \{T | T \text{ is a subfield of } \mathbb{C}, \alpha \in T\}.$$

The *degree* of the number field $\mathbb{Q}(\alpha)$ is the dimension of $\mathbb{Q}(\alpha)$ as a vector space over $\mathbb{Q}$. If α is an algebraic number of degree n, then

$$\mathbb{Q}(\alpha) = \{a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1} | a_i \in \mathbb{Q}\}.$$

Similarly, we define the number field $\mathbb{Q}(\alpha_1, \ldots, \alpha_n)$ for $\alpha_1, \ldots, \alpha_n \in \mathcal{A}$ as $\mathbb{Q}(\alpha_1, \ldots, \alpha_n) := \bigcap\{T|T \text{ is a subfield of } \mathbb{C}, \alpha_1, \ldots, \alpha_n \in T\}$. It holds that for every $\alpha_1, \ldots, \alpha_n \in \mathcal{A}$, there exists $\gamma \in \mathcal{A}$ such that $\mathbb{Q}(\alpha_1, \ldots, \alpha_n) = \mathbb{Q}(\gamma)$.

A number $\beta \in \mathbb{C}$ is called *an algebraic integer* if there is a monic polynomial $f \in \mathbb{Z}[x]$ such that $f(\beta) = 0$. The set of all algebraic integers is denoted by $\mathbb{B}$. *The ring of integers* of the number field $\mathbb{Q}(\alpha)$ is the set $\mathcal{O}_{\mathbb{Q}(\alpha)} := \mathbb{Q}(\alpha) \cap \mathbb{B}$.

Let $s : \mathbb{Q}(\alpha) \to \mathbb{Q}(\alpha)$ be a linear transformation. Moreover, let $S^B \in \mathbb{Q}^{n,n}$ be the matrix of the transformation s in a basis B. It holds that if S^{B_1} and S^{B_2} are two matrices of the same transformations but in different bases, then S^{B_1} is similar to S^{B_2} (i.e., there exists an invertible matrix U such that $S^{B_1} = US^{B_2}U^{-1}$). Especially, they have the same determinant. This means that we can define *the determinant* of the transformation s as $\det(s) = \det(S)$, where S is an arbitrary matrix of the transformation s.

We associate to each element $\delta \in \mathbb{Q}(\alpha)$ a linear transformation $t_\delta : \mathbb{Q}(\alpha) \to \mathbb{Q}(\alpha)$ which is defined by

$$t_\delta(x) = \delta x \tag{1}$$

for every $x \in \mathbb{Q}(\alpha)$.

The matrix of this transformation is denoted T_δ.

Let $\beta \in \mathbb{A}$ and $\gamma \in \mathbb{Q}(\beta)$. Then *the norm* $N_{\mathbb{Q}(\beta)|\mathbb{Q}}(\gamma)$ (or simply $N(\gamma)$ if it is clear in which number field γ lies) of γ is the determinant of the matrix representation of the linear transformation t_γ. In other words

$$N_{\mathbb{Q}(\beta)|\mathbb{Q}}(\gamma) = \det(T_\gamma) \in \mathbb{Q}.$$

A *unit* in a ring R with identity 1_R is an invertible element u of R, i.e., there exists an element $v \in R$ such that $uv = vu = 1_R$. The units of a ring R form a group with respect to multiplication, we call it the *group of units* $U(R)$ of R. In the ring of integers $\mathcal{O}_{\mathbb{Q}(\alpha)}$ of a number field $\mathbb{Q}(\alpha)$, we can characterize the group of units in the following way. Let $\beta \in \mathcal{O}_{\mathbb{Q}(\alpha)}$. Then $\beta \in U(\mathcal{O}_{\mathbb{Q}(\alpha)})$ if and only if $N(\beta) = \pm 1$. Due to the Dirichlet's unit theorem, we can also determine the rank (the number of multiplicatively independent generators) of the group of units $U(\mathcal{O}_{\mathbb{Q}(\alpha)})$.

Theorem 1 (Dirichlet's unit theorem)**.** *Let $K = \mathbb{Q}(\alpha)$ be a number field. The group of units of $\mathcal{O}_K$ is finitely generated and its rank is equal to*

$$r = r_1 + r_2 - 1,$$

where r_1 is the number of real conjugates of α and $2r_2$ is the number of nonreal complex conjugates of α.

For example, if α is a cubic number, then the group of units $U(\mathcal{O}_K)$ has rank either 2 or 1.

Let r be the rank of $U(\mathcal{O}_K)$. The set of units $u_1, \ldots, u_r$ is called *the set of fundamental units* if it is multiplicatively independent and it generates (modulo roots of unity) the group $U(\mathcal{O}_K)$, i.e. if every unit u can be written uniquely in the form

$$u = \zeta u_1^{m_1} \ldots u_r^{m_r}, \tag{2}$$

where $m_i \in \mathbb{Z}$ for all $i \in \{1, \ldots, r\}$ and ζ is some root of unity (i.e. there exists $p \in \mathbb{N}_+$ such that $\zeta^p = 1$).

If $K = \mathbb{Q}(\alpha)$ is an algebraic number field of odd degree, then the roots of unity have the following simple form.

Theorem 2 (Theorem 13.5.2 in [1])**.** *Let $K = \mathbb{Q}(\alpha)$ be an algebraic number field of odd degree. The roots of unity in $\mathcal{O}_K$ are ± 1.*

A number ring with an additive group which is finitely generated is called *an order* in its field of fractions. Let $\alpha_1, \ldots, \alpha_n$ be a basis of the number field $\mathbb{Q}(\alpha_1)$. $\mathbb{Z}[\alpha_1, \ldots, \alpha_n] = \{a_1\alpha_1 + a_2\alpha_2 + \cdots + a_n\alpha_n : a_i \in \mathbb{Z}\} \subset \mathbb{Q}(\alpha)$ is an order of rank n. By [4], the Dirichlet's unit theorem holds also for orders. This means that the group of units of $\mathbb{Z}[\alpha_1, \ldots, \alpha_n]$ is finitely generated by $r = r_1 + r_2 - 1$ generators, where r_1 and r_2 are as in Theorem 1. Moreover, every unit β in $\mathbb{Z}[\alpha_1, \ldots, \alpha_n]$ is also a unit in $\mathcal{O}_{\mathbb{Q}(\alpha)}$ which means that we can find β using the fundamental units of $\mathcal{O}_{\mathbb{Q}(\alpha)}$ and the fact that $N(\beta) = \pm 1$.

2.1 Vectorial MCFs

Let n be a positive integer. A vectorial MCF acts on $\mathbb{R}^n_+$ and it is specified by two sets, $\mathcal{I}$ and $\mathcal{A}$. The first set is a countable set of subsets of $\mathbb{R}^n_+$:

$$\mathcal{I} = \{I_1, I_2, \dots\},$$

while the second set is a set of invertible matrices from $\mathbb{R}^{n,n}$:

$$\mathcal{A} = \{A_1, A_2, \dots\}$$

having the same cardinality as $\mathcal{I}$. Given these two sets, a representation of a vector $\vec{v} \in \mathbb{R}^n_+$ is obtained by the following algorithm.

Algorithm 3 (Multidimensional continued fraction algorithm)**.** *Let $\vec{v} \in \mathbb{R}^n_+$.*

Set $\vec{v}^{(0)} := \vec{v}, i := 0$.

Repeat:

Let j be some index such that $\vec{v}^{(i)} \in I_j$. If there is no such j, the algorithm stops. Otherwise set

$$\vec{v}^{(i+1)} := A_j^{-1}\vec{v}^{(i)}$$

and $A^{(i)} := A_j$. Set $i := i + 1$.

Definition 4. *The sequence $(A^{(i)})_{i=0}^{\infty}$ from Algorithm 3 is called* an $(\mathcal{I}, \mathcal{A})$ $(n-1)$-dimensional continued fraction expansion *of the vector $\vec{v}$.*

If not ambiguous, we will often say only expansion of $\vec{v}$. Moreover, we identify the expansion of $\vec{v}$ with $\vec{v}$, i.e., we write $\vec{v} = (A^{(0)}, A^{(1)}, \dots)$.

A MCF algorithm is *unimodular* if the matrices from $\mathcal{A}$ are unimodular, that is, they have determinant equal to ±1 and integer entries.

An expansion of a vector $\vec{v} = (A^{(0)}, A^{(1)}, \dots)$ is *eventually periodic* if there exists N and positive p such that $A^{(i)} = A^{(i+p)}$ for all $i \geq N$. We write also

$$\vec{v} = \left(A^{(0)}, A^{(1)}, \dots, A^{(N-1)}, \overline{A^{(N)}, A^{(N+1)}, \dots, A^{(N+p-1)}}\right).$$

If $N = 0$, then the expansion is *purely periodic.*

The sequence $\left(A^{(0)}, A^{(1)}, \dots, A^{(N-1)}\right)$ is called a *preperiodic part* and the sequence $\left(A^{(N)}, A^{(N+1)}, \dots, A^{(N+p-1)}\right)$ is called a *repetend.* The number N is called a *preperiod* and the number p is called a *period.*

It follows from Algorithm 3 that

$$A^{(0)} \cdots A^{(i-2)} A^{(i-1)} \vec{v}^{(i)} = \vec{v}^{(0)},$$

we shall consider the preperiodic part and the repetend as matrices, i.e., $R = A^{(0)}A^{(1)} \cdots A^{(N-1)}$ and $M = A^{(N)}A^{(N+1)} \cdots A^{(N+p-1)}$. As a shorthand, we shall use the following notation $\vec{v} = R\overline{M}$.

Below, when we mention a MCF algorithm, we mean an MCF algorithm for some given $(\mathcal{I}, \mathcal{A})$ and n.

2.2 Properties of MCFs

One of the key properties of an MCF algorithm is its convergence, which we will describe next.

2.2.1 Weak convergence

Definition 5. *Let* $\left(M^{(s)}\right)_{s=0}^{+\infty}$ *be a sequence of matrices from* $\mathbb{R}^{n,n}$. *Moreover, let* $j \in \{1, \dots, n\}$. *We say it* weakly converges to $\vec{v} \in \mathbb{R}^n$ with respect to j-th column *if the following two conditions are fulfilled:*

1. *there exists* $\widetilde{P}$ *such that* $M^{(P)}$ *is positive for all* $P > \widetilde{P}$*;*
2. *the sequence*

$$\left(\frac{M_{i,j}^{(s)}}{M_{k,j}^{(s)}}\right)_{s=P}^{+\infty}$$

converges to $\frac{\vec{v}_i}{\vec{v}_k}$ *for all* $i \in \{1, \dots, n\}$ *and some* $k \in \{1, \dots, n\}$.

Remark 6. All elements of all matrices $M^{(s)}$ for $s \geq \widetilde{P}$ are positive and therefore we can choose k arbitrarily.

The $(\mathcal{I}, \mathcal{A})$ $(n-1)$-dimensional MCF algorithm is *weakly convergent* if for every vector $\vec{v} \in \mathbb{R}^n_+$ whose expansion is $\left(A^{(0)}, A^{(1)}, \dots\right)$ with $M^{(s)} = A^{(0)}A^{(1)} \dots A^{(s)}$ we have that the sequence $M^{(s)}$ weakly converges to $\vec{v}$ with respect to the j-th column for every j.

2.2.2 Periodicity of MCFs

The importance of periodicity can be seen from the following theorem.

Theorem 7 ([2], Theorem 3.1.). *Let* $\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{R}^n_+$, $\vec{v} = \overline{M}$ *in a given unimodular MCF algorithm and* M *a be matrix of a repetend of* $\vec{v}$. *We have*

$$\vec{v} = \lambda M \vec{v},$$

where $\lambda \in \mathbb{R}$ *and:*

- λ *is an algebraic unit of degree at most* n.
- *If the degree of* λ *equals* n, *then the numbers* $\frac{v_1}{v_n}, \dots, \frac{v_{n-1}}{v_n}, \frac{v_n}{v_n}$ *constitute a basis (as a vector space over* $\mathbb{Q}$*) of the number field* $\mathbb{Q}(\lambda)$.

The proof of this theorem is based on the fact that $\vec{y}$ is an eigenvector of M and λ^{-1} is the corresponding eigenvalue.

We cannot omit the condition on the degree of λ since $\deg(\lambda) \leq n-1$ would allow $\frac{v_j}{v_n} \notin \mathbb{Q}(\lambda)$. For an example of such a vector and algorithm see Remark (1) in [2].

3 Weakly convergent sequences

For now, we focus on weakly convergent sequences converging to a basis of a number field $\mathbb{Q}(\alpha)$ (as a vector space over $\mathbb{Q}$) of degree n. In the next theorem we show that if these matrices are matrices of multiplication by an element of the field, any matrix of the sequence can be reconstructed from one of its columns.

Theorem 8. *Let* $\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ *be a basis (of a finite field extension of degree n as a vector space over $\mathbb{Q}$), $\ell \in \{1, \dots, n\}$ and $\lambda_s \in \mathbb{Q}(v_1, \dots, v_n)$. There exists a mapping $\mathcal{Q}_{\ell,\vec{v}} : \mathbb{R}^n \mapsto \mathbb{R}^{n,n}$ such that for all sequences $\left(M^{(s)}\right)_{s=0}^{+\infty}$ satisfying*

1. *for all s, $M^{(s)} = {T^{\vec{v}}_{\lambda_s}}^T$, i.e., $M^{(s)}$ equals the transposed matrix of the linear transformation t_{λ_s} in the basis $\vec{v}$ with $\lambda_s \in \mathbb{Q}(v_1, \dots, v_n)$;*
2. *$\left(M^{(s)}\right)_{s=0}^{+\infty}$ weakly converges to $\vec{v}$ with respect to ℓ-th column;*

we have

$$M^{(s)} = \mathcal{Q}_{\ell,\vec{v}}\left(M^{(s)}_{\bullet,\ell}\right) \qquad \text{for any } s.$$

Moreover, there exists an n-tuple $Q_{\ell,\vec{v}}$ of matrices from $\mathbb{Q}^{n,n}$ such that its i-th component satisfies

$$\left(Q_{\ell,\vec{v}}\right)_i M^{(s)}_{\bullet,\ell} = \left(\mathcal{Q}_{\ell,\vec{v}}\left(M^{(s)}_{\bullet,\ell}\right)\right)_{\bullet,i}.$$

In what follows, we keep the same notation as in Theorem 8, i.e., we associate with the mapping $\mathcal{Q}_{\ell,\vec{v}}$ the n-tuple of matrices $Q_{\ell,\vec{v}}$.

4 Periodic MCF expansions

In this section, we will use the results on weakly convergent sequences of matrices from the previous section to investigate periodic MCF expansions. While considering a periodic expansion of $\vec{v}$, i.e., while having $\vec{v} = R\overline{M}$, where the periodic part is already represented as a product M of the matrices of the periodic part of the expansion, we will not distinguish between purely and eventually periodic sequences by considering the matrix RMR^{-1}, called the *matrix of repetend*, and the equality $R\overline{M} = \overline{RMR^{-1}}$. Doing that, we transform the question of finding (if possible) the matrices R and M to finding the decomposition of a candidate matrix Q into the form RMR^{-1}, where $R = R_1 \dots R_k$, $M = M_1 \dots M_\ell$ and $(R_1, \dots, R_k, \overline{M_1, \dots, M_\ell})$ is a MCF expansion of $\vec{v}$.

The following theorem states that the matrix of repetend always equals to a matrix of multiplication by some unit.

Theorem 9. *Let* $\vec{v} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ *be a basis of $\mathbb{Q}(y_1)$ (as a vector space over $\mathbb{Q}$), where $\vec{v}$ has a periodic expansion in a unimodular MCF algorithm. Moreover, let M be a matrix of* *repetend of this MCF expansion of $\vec{v}$. We have*

$$M = T_{\varepsilon}^{\vec{v}T},$$

where $\varepsilon \in U(\mathbb{Z}[\beta y_1, \ldots, \beta y_n])$, $\beta \in \mathbb{Z}, \beta \neq 0$ is such that $\beta y_1, \ldots, \beta y_n$ are algebraic integers, and $T_{\varepsilon}^{\vec{v}}$ is a matrix of linear transformation t_{ε} (defined by (1)*) in the basis $\vec{v}$.*

Remark 10. Let $\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ be a basis of a number field of degree n (as a vector space over $\mathbb{Q}$), $\beta \in \mathbb{Z}, \beta \neq 0$ be such that $\beta y_1, \ldots, \beta y_n$ are algebraic integers, $\varepsilon, \widehat{\varepsilon} \in U(\mathbb{Z}[\beta v_1, \ldots, \beta v_n])$ and $m \in \mathbb{Z}$. Then

$$T_{\varepsilon}^{\vec{v}} T_{\widehat{\varepsilon}}^{\vec{v}} = T_{\widehat{\varepsilon}}^{\vec{v}} T_{\varepsilon}^{\vec{v}} \text{ and } T_{\varepsilon^m}^{\vec{v}} = \left(T_{\varepsilon}^{\vec{v}}\right)^m. \tag{3}$$

this implies that

$$M, N \in \{T_{\varepsilon}^{\vec{v}T} | \varepsilon \in U(\mathbb{Z}[\beta v_1, \ldots, \beta v_n])\} \implies MN = NM.$$

Corollary 11. *Let y be an algebraic number of degree n with minimal polynomial equal to*

$$\sum_{j=0}^{n-1} \alpha_j y^j + y^n = 0,$$

where $\alpha_j \in \mathbb{Q}$ and $\alpha_0 > 0$.

The vector $\vec{v} = \begin{pmatrix} y^{n-1} \\ \vdots \\ y \\ 1 \end{pmatrix}$ does not have a purely periodic expansion in any weakly-convergent $(n-1)$-dimensional continued fraction algorithm for which $\mathcal{A} \subset \mathrm{SL(n, \mathbb{N})}$.

4.1 Finding candidates on the matrix of repetend

Lemma 12. *Let $\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ be a basis of $\mathbb{Q}(v_1)$ as a vector space over $\mathbb{Q}$. We can determine the n-tuples $Q_{k,\vec{v}}$ (for $k \in \{1, \ldots, n\}$) directly from the first n powers of the matrix M of the repetend of the vector $\vec{v}$ (in some unimodular wakly-convergent MCF algorithm).*

Based on this idea from the last lemma, we show in this section how to find matrices that could potentially be the matrices of repetend of an MCF expansion of $\vec{v}$ in a given weakly convergent MCF algorithm. We call such a matrix a *candidate on the matrix of repetend.*

For the sake of simplicity, we do this explicitly for $n = 3$. We start with a lemma which is an explicit version of **??**.

Lemma 13. *Let y be a cubic number with minimal polynomial equal to $\alpha_0+\alpha_1 y+\alpha_2 y^2+y^3=0$, where $\alpha_0,\alpha_1,\alpha_2\in\mathbb{Q}$, $x=\gamma_0+\gamma_1 y+\gamma_2 y^2$, where $\gamma_0,\gamma_1,\gamma_2\in\mathbb{Q}$ and $\vec{v}=\begin{pmatrix}x\\y\\1\end{pmatrix}$.*

We have $Q_{1,\vec{v}}=\left(\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\begin{pmatrix}0&b_1&c_1\\1&b_2&c_2\\0&b_3&c_3\end{pmatrix},\begin{pmatrix}0&c_1&c_4\\0&c_2&c_5\\1&c_3&c_6\end{pmatrix}\right)$, *where*

$$b_3=\gamma_2, c_3-b_2=\gamma_1, c_2=-\gamma_0, \alpha_2=\frac{2c_3-b_2}{b_3}, \alpha_1=\frac{c_3^2-b_1-b_3c_2-b_2c_3}{b_3^2}, \alpha_0=\frac{-c_1-c_3c_2}{b_3^2}. \tag{4}$$

and

$$c_4=\frac{c_1c_3-c_1b_2+c_2b_1}{b_3},\quad c_5=\frac{c_3c_2+c_1}{b_3}\quad \textit{and}\quad c_6=\frac{c_3^2+b_3c_2-b_1-b_2c_3}{b_3}. \tag{5}$$

Or equivalently

$$\begin{aligned}
b_1 &= \gamma_2\gamma_0+\gamma_1\alpha_2\gamma_2-\gamma_1^2-\alpha_1\gamma_2^2\\
b_2 &= \alpha_2\gamma_2-2\gamma_1\\
b_3 &= \gamma_2\\
c_1 &= \gamma_0\alpha_2\gamma_2-\gamma_0\gamma_1-\alpha_0\gamma_2^2\\
c_2 &= -\gamma_0\\
c_3 &= \alpha_2\gamma_2-\gamma_1\\
c_4 &= \gamma_0\alpha_1\gamma_2-\gamma_0^2-\alpha_0\gamma_2\gamma_1\\
c_5 &= -\alpha_0\gamma_2\\
c_6 &= \alpha_1\gamma_2-2\gamma_0.
\end{aligned} \tag{6}$$

Remark 14. Let $\widehat{y}$ be a cubic number with minimal polynomial equal to $\alpha_0+\alpha_1\widehat{y}+\alpha_2\widehat{y}^2+\widehat{y}^3=0$, where $\alpha_0,\alpha_1,\alpha_2\in\mathbb{Q}$, $\widehat{x}=\gamma_0+\gamma_1\widehat{y}+\gamma_2\widehat{y}^2$, where $\gamma_0,\gamma_1,\gamma_2\in\mathbb{Q}$ and $\vec{v}=\begin{pmatrix}\widehat{x}\\\widehat{y}\\1\end{pmatrix}$ be a basis of some cubic number field (as a vector space over $\mathbb{Q}$). We find all the candidates on the matrix of repetend of the MCF expansion of the vector $\vec{v}$.

Let $\beta\in\mathbb{Z}$ be such that $\beta\widehat{x},\beta\widehat{y}$ are algebraic integers. Firstly, we have to realise that the number $\beta\widehat{y}$ is a cubic number, and therefore, by Dirichlet's theorem, there are either one or two fundamental units in $\mathbb{Z}[\beta\widehat{y},\beta\widehat{x}]$.

Let $\varepsilon_1=\beta_1+\beta_2\widehat{y}+\beta_3\widehat{x}$ (resp. $\varepsilon_1=\beta_1+\beta_2\widehat{y}+\beta_3\widehat{x}, \varepsilon_2=\widehat{\beta}_1+\widehat{\beta}_2\widehat{y}+\widehat{\beta}_3\widehat{x}$) be the fundamental unit (resp. units) of $\mathbb{Z}[\beta\widehat{y},\beta\widehat{x}]$. (It follows that β is a divisor of $\beta_2,\beta_3,\widehat{\beta}_2,\widehat{\beta}_3$.)

It follows from Theorem 9 and Remark 10 that every candidate M on the matrix of repetend of the MCF expansion of $\begin{pmatrix}\widehat{x}\\\widehat{y}\\1\end{pmatrix}$ can be written as

$$M=\pm\left(T_{\varepsilon_1}^{\vec{v}\,T}\right)^{m_1} \tag{7}$$

for $m_1 \in \mathbb{Z}$ (respectively

$$M = \pm\left(\left({T_{\varepsilon_1}^{\vec{v}}}^T\right)^{m_1}\left({T_{\varepsilon_2}^{\vec{v}}}^T\right)^{m_2}\right) \tag{8}$$

for $m_1, m_2 \in \mathbb{Z}$).

We can easily verify by direct computation that $\left({T_{\varepsilon_1}^{\vec{v}}}^T\right)_{\bullet,1} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix}$, where $x_1 = \beta_1 + \beta_2\left(\frac{\gamma_1}{\gamma_2} - \alpha_2\right) + \beta_3\left(\frac{\gamma_1^2}{\gamma_2} - 2\gamma_1\alpha_2 + \alpha_2^2\gamma_2 + 2\gamma_0 - \alpha_1\gamma_2\right)$, $y_1 = \frac{\beta_2}{\gamma_2} + \beta_3\left(\frac{\gamma_1}{\gamma_2} - \alpha_2\right)$, $z_1 = \beta_3$, and eventually $\left({T_{\varepsilon_2}^{\vec{v}}}^T\right)_{\bullet,1} = \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix}$ where

$$x_2 = \widehat{\beta}_1 + \widehat{\beta}_2\left(\frac{\gamma_1}{\gamma_2} - \alpha_2\right) + \widehat{\beta}_3\left(\frac{\gamma_1^2}{\gamma_2} - 2\gamma_1\alpha_2 + \alpha_2^2\gamma_2 + 2\gamma_0 - \alpha_1\gamma_2\right),$$

$y_2 = \frac{\widehat{\beta}_2}{\gamma_2} + \widehat{\beta}_3\left(\frac{\gamma_1}{\gamma_2} - \alpha_2\right)$, $z_2 = \widehat{\beta}_3$.

Now, we can use Theorem 8 and Lemma 13 to compute the matrices ${T_{\varepsilon_1}^{\vec{v}}}^T$ (resp. ${T_{\varepsilon_2}^{\vec{v}}}^T$). We use the notation from Lemma 13. We get that

$${T_{\varepsilon_1}^{\vec{v}}}^T = \left((Q_{1,\vec{v}})_1 \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} \quad (Q_{1,\vec{v}})_2 \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} \quad (Q_{1,\vec{v}})_3 \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} \right)$$

and similarly for the matrix ${T_{\varepsilon_2}^{\vec{v}}}^T$.

For simplicity, we do the explicit calculation only for the case $\widehat{x} = \widehat{y}^2$. In this case we get a simpler form and that is $x_1 = \beta_1 - \beta_3\alpha_1 - \beta_2\alpha_2 + \beta_3\alpha_2^2, y_1 = \beta_2 - \beta_3\alpha_2, z_1 = \beta_3$ and eventually $x_2 = \widehat{\beta}_1 - \widehat{\beta}_3\alpha_1 - \widehat{\beta}_2\alpha_2 + \widehat{\beta}_3\alpha_2^2, y_2 = \widehat{\beta}_2 - \widehat{\beta}_3\alpha_2, z_2 = \widehat{\beta}_3$.

For $i \in \{1, 2\}$, we get that

$${T_{\varepsilon_i}^{\vec{v}}}^T = \begin{pmatrix} x_i & -\alpha_1 y_i - \alpha_0 z_i & -\alpha_0 y_i \\ y_i & x_i + \alpha_2 y_i & -\alpha_0 z_i \\ z_i & y_i + \alpha_2 z_i & x_i + \alpha_2 y_i + \alpha_1 z_i \end{pmatrix}.$$

Not all of these matrices are the candidates on the matrix of repetend. In fact, only half of them have the determinant equal to 1 (the other half has the determinant equal to −1). We always have to check the determinant of M_{x_1,y_1,z_1} and M_{x_2,y_2,z_2} and choose the sign in correspondence with this determinant. It can also happen that some of these matrices are not integer matrices, and therefore they are not candidates on the matrix of repetend. On the other hand, if $\widehat{y}$ is an algebraic integer, then this problem cannot occur, and therefore we know that every such matrix (with the correct sign of the determinant) is a candidate on the matrix of repetend.

This means that the problem of determining the candidates on the matrix of repetend of the MCF expansion of $\begin{pmatrix} \widehat{x} \\ \widehat{y} \\ 1 \end{pmatrix}$ can be solved by determining units in $\mathbb{Z}[\beta\widehat{y}, \beta\widehat{x}]$.

We illustrate Remark 14 on an example. In this example, we also show the reason, why we beleive that most of the well-known MCF algorithms seems to fail in solving the Hermites question.

Example 15. Let y_1, y_2, y_3 be the three positive real roots of the polynomial $y^3 - 6y^2 + 9y - 3 = 0$. We investigate the MCF expansion of the vectors $\vec{v_1} = \begin{pmatrix} {y_1}^2 \\ y_1 \\ 1 \end{pmatrix}$, $\vec{v_1} = \begin{pmatrix} {y_2}^2 \\ y_2 \\ 1 \end{pmatrix}$ and $\vec{v_3} = \begin{pmatrix} {y_3}^2 \\ y_3 \\ 1 \end{pmatrix}$. The numbers y_1, y_2, y_3 are three real conjugates and therefore there are two fundamental units in $\mathbb{Z}[y_1, y_1^2]$, $\mathbb{Z}[y_2, y_2^2]$ and $\mathbb{Z}[y_3, y_3^2]$. The couples of fundamental units are $\varepsilon_i = y_i^2 - 5y_i + 5, \widetilde{\varepsilon_i} = y_i^2 - 5y_i + 4$ for all $i \in \{1, 2, 3\}$.

Using Lemma 13, we get that

$$Q_{1,\vec{v_i}} = \left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -9 & 3 \\ 1 & -6 & 0 \\ 0 & 1 & -6 \end{pmatrix}, \begin{pmatrix} 0 & 3 & 0 \\ 0 & 0 & 3 \\ 1 & -6 & 9 \end{pmatrix} \right)$$

for every $i \in \{1, 2, 3\}$.

Using the computation in Remark 14, we obtain that every candidate M on the matrix of repetend of the vector $\vec{v_1}, \vec{v_2}$ and also $\vec{v_3}$ is defined by

$$M = \pm(M_1^{m_1} M_2^{m_2})$$

where $m_1, m_2 \in \mathbb{Z}$,

$$M_1 = \begin{pmatrix} 2 & -6 & 3 \\ 1 & -4 & 3 \\ 1 & -5 & 5 \end{pmatrix} \quad \text{and} \quad M_2 = \begin{pmatrix} 1 & -6 & 3 \\ 1 & -5 & 3 \\ 1 & -5 & 4 \end{pmatrix}.$$

This mean, that the candidates on the matrix of repetend are identical for the three vectors. In fact, if we work only with the matrices of repetend and the triplets $Q_{\bullet,\vec{v}}$, we can not distinquish, with which of the three vectors we work. On the other hand (as we will se bellow), the MCF expansions in the well-known algorithms (for example the Brun algorithm) of these three vectors differ. This means, that these well-known algorithms use some information, that is not included in the matrices of repetend. We suppose, that this is the reason, why these algorithms seem to fail to answer the Hermite's question.

We compare it with the expansion of $\vec{v_1}, \vec{v_2}$ and $\vec{v_3}$ in the Brun and in the Selmer algorithm.

In the Brun algorithm, the vector $\vec{v_1}, \vec{v_2}, \vec{v_3}$ have the following periodic expansions. The vector $\vec{v_1} = T_{32}, \overline{T_{32}T_{21}^2T_{13}^3T_{32}^2T_{21}T_{13}^6} = \overline{M_{B,1}}$ where

$$M_{B,1} = \begin{pmatrix} 7 & -39 & 45 \\ 15 & -83 & 96 \\ 32 & -177 & 205 \end{pmatrix} = M_1^4 M_2.$$

The vector $\vec{v_2} = T_{12}T_{21}T_{13}T_{32}, \overline{T_{23}T_{32}^2T_{23}T_{31}T_{13}T_{31}T_{13}T_{32}T_{23}^2T_{32}T_{21}T_{12}T_{21}T_{12}} = \overline{M_{B,2}}$ where

$$M_{B,2} = \begin{pmatrix} -590 & 2565 & -1071 \\ -357 & 1552 & -648 \\ -216 & 939 & -392 \end{pmatrix} = M_1^{-6} M_2^6.$$

And finally the vector $\vec{v_3} = T_{12}^3, \overline{T_{21}T_{13}^3T_{32}^2T_{21}T_{13}^6T_{32}T_{21}} = \overline{M_{B,3}}$ where

$$M_{B,3} = -\begin{pmatrix} 256 & -543 & 198 \\ 66 & -140 & 51 \\ 17 & -36 & 13 \end{pmatrix} = M_1 M_2^{-5}.$$

In the Selmer algorithm, we did not find a periodicpart for the vectors $\vec{v_1}, \vec{v_2}, \vec{v_3}$ in the first 10000 steps.

5 Conclusion

In this text we showed a new approach to the problem of periodicity of MCF algorithms. In Theorem 9, we proved that ever matrix of repetend of an expansion in a MCF algorithm is equal to a matrix of some linear transformation. Moreover, we provided tools, how to find these matrices.

Putting these informations together, we showed that there exist some vectors that can not have a purely periodic expansion in any unimodular weakly convergent MCF algorithm. We also provided an example which shows a problem which could be the cause why most of the well-known MCF algorithms seems to fail to answer the Hermite quesiton.

References

[1] Ş. Alaca and K. S. Williams. *Introductory algebraic number theory.* Cambridge University Press Cambridge, (2004).

[2] A. J. Brentjes. *Multi-dimensional continued fraction algorithms.* MC Tracts (1981).

[3] V. Brun. *En generalisation av Kjedebroken.* na, (1920).

[4] K. Conrad. Dirichlet's unit theorem.

[5] E. Heine and C. Jacobi. *Allgemeine Theorie der kettenbruchähnlichen Algorithmen, in welchen jede Zahl aus drei vorhergehenden gebildet wird.* Journal für die reine und Angewandte Mathematik **69** (1868), 29–64.

[6] C. Hermite. *Letter to CDJ Jacobi.* J. Reine Angew. Math **40** (1839), 286.

[7] O. Karpenkov. *Geometry of continued fractions*, volume 26. Springer Science & Business Media, (2013).

[8] S. Labbé. 3*-dimensional continued fraction algorithms cheat sheets.* arXiv preprint arXiv:1511.08399 (2015).

[9] O. Perron. *Ein Satz über Jacobi-Ketten zweiter Ordnung.* Annali della Scuola Normale Superiore di Pisa-Classe di Scienze **4** (1935), 133–138.

[10] H. Poincaré. *Sur une généralisation des fractions continues.* CR Acad. Sci. Paris. Ser **1** (1884), 1014–1016.

[11] F. Schweiger. *Invariant measures for fully subtractive algorithms.* Anz. Österreich. Akad. Wiss. Math.-Natur. Kl **131** (1995), 25–30.

[12] F. Schweiger. *Multidimensional continued fractions.* Oxford University Press on Demand, (2000).

[13] E. S. SELMER. *Om flerdimensjonal kjedebrøk.* Nordisk Matematisk Tidskrift (1961), 37–43.

[14] J.-i. Tamura and S.-i. Yasutomi. *A new multidimensional continued fraction algorithm.* Mathematics of computation **78** (2009), 2209–2222.

Quantum Walk Search and State Transfer with Fully Connected Vertices[*]

Stanislav Skoupý

3rd year of PGS, email: Stanislav.Skoupy@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. We examine search algorithm and state transfer algorithm based on discrete-time quantum walks with coins and how they perform on graphs with fully connected vertices which are vertices connected to all other vertices in the graph. We show that search algorithm does not always find marked fully connected vertex with probability close to one. But we prove that the state transfer algorithm achieves to move walker between two fully connected vertices with probability close to one for all graphs. We also prove that search for fully connected vertex can be improved by changing the initial state of the algorithm.

Keywords: quantum walk, search algorithm, state transfer algorithm

Abstrakt. Zkoumáme vyhledávací algoritmus a algoritmus pro přenos stavu založených na kvantové procházce v diskrétním čase s mincí a jak fungují na grafu s plně napojenými vrcholy, což jsou vrcholy, které jsou připojeny ke všem ostatním vrcholům grafu. Ukážeme, že vyhledávací algoritmus nenajde vždy s pravděpodobností blízkou jedné plně napojený označený vrchol. Ale dokážeme, že algoritmus pro přenos stavu přenese chodce mezi dvěma plně napojenými vrcholy s pravděpodobností blízkou jedné pro všechny grafy. Také dokážeme, že vyhledávání plně napojeného vrcholu může být vylepšeno změnou počátečního stavu algoritmu.

Klíčová slova: kvantová procházka, vyhledávací algoritmus, algoritmus na přesnos stavu

1 Introduction

Quantum algorithms based on quantum walks are important part of quantum computer science. In this article we examine the search and state transfer algorithm based on discrete-time quantum walks with coins on graphs that contains fully connected vertices which are vertices connected to all other vertices in the graph.

We follow the work of [9] where they prove that using Laplacian matrix as a Hamiltonian on graph with fully connected vertex leads to finding such a vertex with probability close to one. Laplacian matrix is defined as $L = D - A$ where A is adjacency matrix and D is diagonal matrix with degree of vertices d_v as a diagonal elements, i.e. $D_{v,v} = d_v$. In this work we examine the same graphs using model of discrete-time quantum walks with coins.

[*]This work was supported from Student Grant Competition of Czech Technical University in Prague under Grant SGS19/186/OHK4/3T/14 and project CAAS.

In the first part we introduce general search and state transfer algorithm based on discrete-time quantum walks and in second part we examine those algorithms on the graphs with fully connected vertices.

2 General search and state transfer algorithm

In this section we present model of discrete time-quantum walks with coins and search and state transfer algorithm which are based on them. The search algorithm (SA) was introduce in following articles [2] and [3] and state transfer algorithm (STA) is based on work in [4].

Starting with the concept of discrete-time quantum walks let us have a graph $G = (V, E)$, then we assign to every vertex v subspace $\mathcal{H}_v$. $\mathcal{H}_v$ is known as coin-space corresponding to vertex v and it reads $\mathcal{H}_v = \text{Span}\left\{|v,w\rangle, w \in V : \{v,w\} \in E\right\}$, where the first index v describe the position of the walker and the second index describe w describes the direction of the walker. The Hilbert space of the graph G is a direct sum of all coin spaces $\mathcal{H}_G = \bigoplus_v \mathcal{H}_v$. 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 the shift operator in the evolution of the walk is too simple since $\hat{S}^2 = \hat{I}$. Hence, the so-called coin operator $\hat{C}$ is additionally applied at every step. The coin operator acts locally at each subspace $\mathcal{H}_v$. So the coin operator has a form $\hat{C} = \bigoplus_v \hat{C}_v^{(l)}$, where $\hat{C}_v^{(l)}$ is local unitary coin operator. Hence, 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. $\hat{U} = \hat{S}\hat{C}$.

Now that we introduce basic notation and concept of the discrete time quantum walk we move to introduction of SA and STA, starting with SA. The main idea of SA is applying one local coin operator to marked vertices and different local coin operator to other non-marked vertices of the graph. The local coin operator that we use at non-marked vertices is known as Grover operator [5]

$$\hat{G}_v^l = -\hat{I} + 2|\Omega_v\rangle\langle\Omega_v| \tag{2}$$

where $|\Omega_v\rangle$ is equal superposition of all direction of the coin space at vertex v. $\hat{G}_v^l$ operator acts locally at $\mathcal{H}_v$. At the marked vertices often used coins are simple phase shift by π or the Grover operator followed by phase shift by π. In this article we use modified Grover coin with additional weight on the state corresponding to the loop at marked vertex followed by phase shift by π. In [6] and [7] Wong showed that adding additional loop at each vertex and properly tuning the weight of the loop the improves probability of finding the marked vertex at d-regular graph. To add this weighted loop we modified the Grover operator by replacing state $|\Omega_v\rangle$ with state given by

$$|\Omega_v(l)\rangle = \frac{1}{\sqrt{d_v + l}} \left(\sum_{w,\{v,w\}\in E} |v,w\rangle + \sqrt{l}|v,v\rangle \right) \tag{3}$$

where d_v is number of neighbours of vertex v, state $|v,v\rangle$ corresponds to loop and l is real number corresponding to the weight of the loop. Using state $|\Omega_v(l)\rangle$ we get modified Grover operator $\hat{G}_v^{(l)}(l) = -\hat{I} + 2|\Omega_v(l)\rangle\langle\Omega_v(l)|$. We label the choice of marked coin as $\hat{C}_m^{(l)}$ for marked vertex m. Then the global coin operator of the search algorithm reads

$$\hat{C}_m = \bigoplus_v \hat{G}_v^{(l)} \oplus \hat{C}_m^{(l)} \tag{4}$$

Using coins operator (4) 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

$$|\Omega\rangle = \frac{1}{\sqrt{2|E|}} \sum_{v\in V} \sum_{\substack{w \\ \{v,w\}\in E}} |v,w\rangle. \tag{5}$$

2. Apply the evolution operator $\hat{U}_m = \hat{S}\hat{C}_m$ T-times, i.e $|\phi(T)\rangle = \hat{U}_m^T|\Omega\rangle$

3. Measure the system.

The probability of success is given by

$$p_m(T) = \sum_{\substack{w \\ \{m,w\}\in E}} |\langle m,n|\phi\,(T)\rangle|^2. \tag{6}$$

i.e. the probability that the walker is located at marked vertex at the end of the algorithm. The success probability and number of steps T depends on the structure and the size of the graph G.

Having introduced the SA we move to the STA. In the case of STA, we have now two marked vertices between which we want to transfer the walker. Let us called them the sender and the receiver and we label corresponding vertices by s and r, respectively. The coin of STA is now given by

$$\hat{C}_{s,r} = \bigoplus_{\substack{v\in V \\ v\neq s,r}} \hat{G}_v^{(l)} \oplus \hat{C}_s^{(l)} \oplus \hat{C}_r^{(l)}, \tag{7}$$

which we use in the construction of evolution operator as follows $\hat{U}_{s,r} = \hat{S}\hat{C}_{s,r}$. There is also a change of initial state of the walk. The initial state of STA is always localized at sender vertex, equal superposition of all direction at sender vertex $|\Omega_s\rangle$ or state corresponding to loop $|s,s\rangle$ are often chosen as initial state of the algorithm. Success of STA depends on the choice of initial state, so the proper choice has to be done before the run of the algorithm. In this work we use loop state as a initial state of STA. Steps of STA are following:

1. Initialize the system in the state corresponding to a loop at sender vertex $|s,s\rangle$.

2. Apply the evolution operator $\hat{U}_{s,r}$ T-times.

3. Measure the system. The particle moves from the sender to the receiver with fidelity $\mathcal{F}(T)$ given by

$$\mathcal{F}(T) = \sum_{\substack{w \\ \{r,w\}\in E}} |\langle r, w|\psi\,(T)\rangle|^2. \qquad (8)$$

The number of steps T and the fidelity $\mathcal{F}(T)$ depend again on the structure and the size of the graph. We say that the algorithm achieves perfect state transfer if the fidelity is close to 1, i.e. the particle moves from the sender to the receiver with probability close to one.

3 Fully connected vertices in random graph

In this section we examine search algorithm and state transfer algorithm on fully connected vertices. In the case of SA we use model of Erdös-Rényi graph $G(N-1,q)$, which is a graph with $N-1$ vertices, where each edge has probability q to be in the graph, then we add one vertex m to the graph G and we connect it to all other vertices. In the case of STA we use $G(N-2,q)$ and we add vertices of sender and receiver in the same manner. Moreover, we use as non-marked coin modified Grover coin $\hat{G}^l_v(l)$ where we tuned the weight at each vertex $l_v = N - d_v$, i.e. we get $\hat{G}^l_v(N-d_v)$. This choice is done so it corresponds with [9], where they have the weight of the loop of d_v. However if we subtract N times identity operator $N\hat{I}$ from their Hamiltonian we get the weight of the loop $d_v - N$ and evolution of the continuous walk would not change. On marked vertices we use $-\hat{G}^l_v(1)$ as the local coin operator.

Starting from the search for fully connected vertex numerical simulation showed that algorithm finds the marked vertex but not always with probability close to one depending on probability q in $G(N-1,q)$, probability of success sinks with lowering of probability q. For illustration of this effect see Fig. 1 and 2. Decline of success probability holds to certain point of q, because for very small q most of the edges in the graph are edges connected to marked vertex and the initial state has a large overlap with marked vertex. For very small q probability of success starts to rise again up to 0.5.

Graphs where SA finds marked vertex with probability close to one are good very often graphs where STA achieves perfect state transfer [8]. Also when the SA does not find marked vertex with probability STA usually likewise fail to perform perfect state transfer. But numerical simulation suggests that STA achieves perfect state transfer independent of q using the model of 2 fully connected vertices adjacent to $G(N-2,q)$ with $\hat{G}^l_v(N-d_v)$ as non-marked coins. Moreover we proves this analytically. We find the invariant subspace of the walk $\mathcal{I}$ which reads

$$\begin{aligned}
|\nu_1\rangle &= |s,s\rangle & |\nu_2\rangle &= |r,r\rangle \\
|\nu_3\rangle &= |s,r\rangle & |\nu_4\rangle &= |r,s\rangle \\
|\nu_5\rangle &= \tfrac{1}{\sqrt{N-2}} \sum_{v=3}^{N} |s,v\rangle & |\nu_6\rangle &= \tfrac{1}{\sqrt{N-2}} \sum_{v=3}^{N} |r,v\rangle \\
|\nu_7\rangle &= \tfrac{1}{\sqrt{N-2}} \sum_{v=3}^{N} |v,s\rangle & |\nu_8\rangle &= \tfrac{1}{\sqrt{N-2}} \sum_{v=3}^{N} |v,r\rangle
\end{aligned}$$

$$|\nu_9\rangle = \tfrac{1}{N-2}\sum_{v=3}^{N}\Big(\sqrt{N}\,|\Omega_v(N-d_v)\rangle - |v,r\rangle - |v,s\rangle\Big).$$

$\mathcal{I}$ is invariant with to evolution operator $\hat{U}_{s,r}$ and it contains the initial and target state of the STA. Initial state corresponding to a loop at sender is $|\nu_1\rangle$ and fidelity of state transfer reads

$$\mathcal{F}(t) = \sum_{\substack{w\\ \{r,w\}\in E}} |\langle r,w|\psi\,(t)\rangle|^2 = |\langle \nu_2|psi(T)\rangle|^2 + |\langle\nu_4|\psi(T)\rangle|^2 + |\langle\nu_6|\psi(T)\rangle|^2. \tag{9}$$

We introduce the effective evolution operator $\hat{U}_{eff}$ which is evolution operator $\hat{U}_{s,r}$ acting in subspace $\mathcal{I}$ as follows

$$\begin{aligned}
\hat{U}_{eff}\,|\nu_1\rangle &= \frac{1}{N}\Big((N-2)\,|\nu_1\rangle - 2\,|\nu_4\rangle - 2\sqrt{N-2}\,|\nu_7\rangle\Big)\\
\hat{U}_{eff}\,|\nu_2\rangle &= \frac{1}{N}\Big((N-2)\,|\nu_2\rangle - 2\,|\nu_3\rangle - 2\sqrt{N-2}\,|\nu_8\rangle\Big)\\
\hat{U}_{eff}\,|\nu_3\rangle &= \frac{1}{N}\Big(-2\,|\nu_1\rangle + (N-2)\,|\nu_4\rangle - 2\sqrt{N-2}\,|\nu_7\rangle\Big)\\
\hat{U}_{eff}\,|\nu_4\rangle &= \frac{1}{N}\Big(-2\,|\nu_2\rangle + (N-2)\,|\nu_3\rangle - 2\sqrt{N-2}\,|\nu_8\rangle\Big)\\
\hat{U}_{eff}\,|\nu_5\rangle &= \frac{1}{N}\Big(-2\sqrt{N-2}\,|\nu_1\rangle - 2\sqrt{N-2}\,|\nu_4\rangle + (4-N)\,|\nu_7\rangle\Big)\\
\hat{U}_{eff}\,|\nu_6\rangle &= \frac{1}{N}\Big(-2\sqrt{N-2}\,|\nu_2\rangle - 2\sqrt{N-2}\,|\nu_3\rangle + (4-N)\,|\nu_8\rangle\Big)\\
\hat{U}_{eff}\,|\nu_7\rangle &= \frac{1}{N}\Big((2-N)\,|\nu_5\rangle + 2\,|\nu_6\rangle + 2\sqrt{N-2}\,|\nu_9\rangle\Big)\\
\hat{U}_{eff}\,|\nu_8\rangle &= \frac{1}{N}\Big(2\,|\nu_5\rangle + (2-N)\,|\nu_6\rangle + 2\sqrt{N-2}\,|\nu_9\rangle\Big)\\
\hat{U}_{eff}\,|\nu_9\rangle &= \frac{1}{N}\Big(2\sqrt{N-2}\,|\nu_5\rangle + 2\sqrt{N-2}\,|\nu_6\rangle + (N-4)\,|\nu_9\rangle\Big)
\end{aligned}$$

This reduction of dimension where the $\hat{U}_{eff}$ is does not depend on q allow us to calculate the evolution of fidelity of STA which reads in the limit of large graph

$$\mathcal{F}(t) = \sin^4\left(\frac{\omega t}{2}\right) \tag{10}$$

where ω is eigenphase of pair of conjugate eigenvalues which has eigenvectors with large overlap with initial state. The ω is given by

$$\omega = \arccos\left(\sqrt{\frac{N-2}{N}}\right). \tag{11}$$

From (14) we see that STA achieves perfect state transfer after number of steps which is the closet integer to

$$T \approx \frac{\pi}{\omega} = \frac{\pi}{\arccos\left(\sqrt{\frac{N-2}{N}}\right)} \approx \frac{\pi\sqrt{N}}{\sqrt{2}} + O\left(\frac{1}{\sqrt{N}}\right) \tag{12}$$

For comparison of analytical and numerical results see Fig. 3.

Let us now return to SA where we did no found any invariant subspace. The reason for that is that initial state of SA does not have same overlap with each vertex of the graph. The overlap depends on the degree of vertex d_v. We find that if we change the initial state of SA to

$$|\Omega\rangle = \frac{1}{\sqrt{N}} \sum_{v \in V} |\Omega_v(N - d_v)\rangle \,, \tag{13}$$

which has the same overlap with all vertices $\frac{1}{\sqrt{N}}$, there is invariant subspace $\mathcal{I}$ of the walk which is spanned by following states

$$\begin{aligned} |\nu_1\rangle &= |m, m\rangle \\ |\nu_2\rangle &= \frac{1}{\sqrt{N-1}} \sum_{v=2}^{N} |m, v\rangle \\ |\nu_3\rangle &= \frac{1}{\sqrt{N-1}} \sum_{v=2}^{N} |v, m\rangle \\ |\nu_4\rangle &= \frac{1}{N-1} \sum_{v=2}^{N} \left(\sqrt{N}\, |\Omega_v(N - d_v)\rangle - |v, m\rangle\right) . \end{aligned}$$

We again introduce the effective evolution operator of the SA as evolution operator acting in $\mathcal{I}$ which is given by following matrix

$$U_{eff} = \begin{pmatrix} \frac{N-2}{N} & -\frac{2\sqrt{N-1}}{N} & 0 & 0 \\ 0 & 0 & \frac{2-N}{N} & \frac{2\sqrt{N-1}}{N} \\ -\frac{2\sqrt{N-1}}{N} & \frac{2-N}{N} & 0 & 0 \\ 0 & 0 & \frac{2\sqrt{N-1}}{N} & \frac{N-2}{N} \end{pmatrix} .$$

After some calculation we get evolution of success probability which reads

$$\mathcal{F}(t) = \sin^2\left(\frac{\omega t}{2}\right) \tag{14}$$

where ω is again the eigenphase of pair of conjugate eigenvalues with eigenvectors with largest overlap with initial state and it reads

$$\omega = \arccos\left(\frac{N-2}{N}\right) . \tag{15}$$

The number of steps is closest integer to number

$$T \approx \frac{\pi}{\omega} = \frac{\pi}{\arccos\left(\frac{N-2}{N}\right)} \approx \frac{\pi\sqrt{N}}{2} + O\left(\frac{1}{\sqrt{N}}\right) \tag{16}$$

It is easy to see that success probability for SA with new initial state (13) goes to one for graphs independent of q.

4 Conclusion

We show that the SA with original initial state does not finds marked fully connected vertex with probability close to one independent of q which is probability of edge being in Erdös-Rényi graph $G(N-1,q)$. Nevertheless the STA performs perfect state transfer between two fully connected vertices with properly tuned weights of the loops at non-marked vertices. Also we showed that change of initial state is SA leads to succes probability close to one for all graphs.

References

[1] L. Razzoli, P. Bordon, M. G. A. Paris *Universality of the fully connected vertex in Laplacian continuous-time quantum walk problems*, arXiv: 2202.13824 (2022)

[2] N. Shenvi, J. Kempe, K. B. Whaley. *A quantum random walk search algorithm*, Phys. Rev. A 67, 052307 (2003)

[3] 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

[4] B. Hein, G. Tanner. *Wave communication across regular lattice*, Phys. Rev. Lett. 103, 260501 (2009)

[5] Lov K. Grover. *Quantum Mechanic helps in searching for a needle in a haystack*, Phys. Rev Lett. 78, 325 (1997)

[6] T. G. Wong. *Grover Search with Lackadaisical Quantum Walks*, J.Phys. A 48, 435304 (2015)

[7] T. G. Wong. *Coined quantum walks on weighted graphs*, Journal of Physics A: Mathematical and Theoretical 50, (2017)

[8] D. Reitzner, M. Hillery, E. Feldman, V. Bužek *Quantum searches on highly symmetric graphs*, Phys. Rev. A 79, 012323 (2009)

[9] L. Razzoli, P. Bordon, M. G. A. Paris *Universality of the fully connected vertex in Laplacian continuous-time quantum walk problems*, arXiv: 2202.13824 (2022)

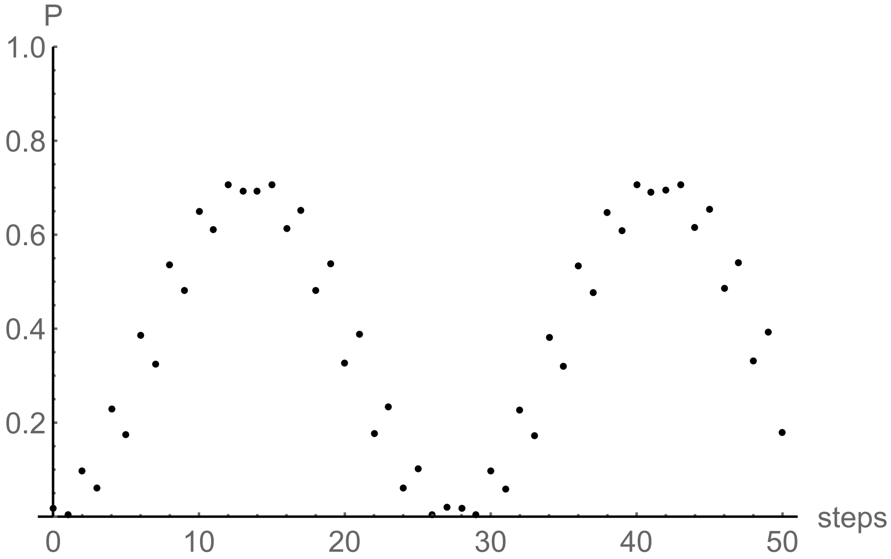


Figure 1: Evolution of success probability during the run of search algorithm on graph with fully connected vertex connected to graph $G(79, 0.7)$

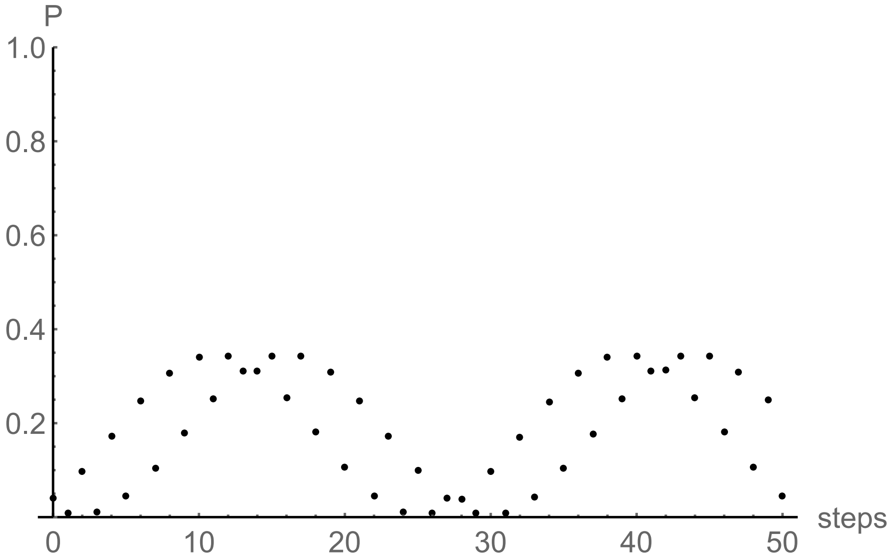


Figure 2: Evolution of success probability during the run of search algorithm on graph with fully connected vertex connected to graph $G(79, 0.3)$

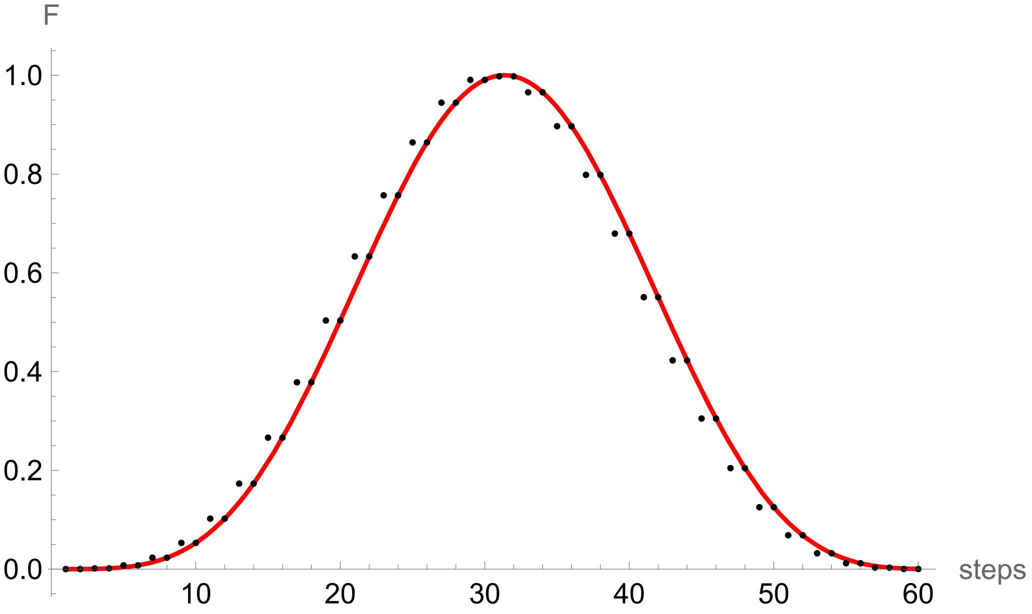


Figure 3: Evolution of fidelity during the run of STA on graph between 2 connected vertices on a graph with 200 vertices. Line is analytical result (14), dots are numerical simulation. Difference between analytical and numerical result at odd steps is due to the limit of large graph for analytical result.

Combining Machine Learning and Mathematical Modeling in Estimation of T_1 Relaxation Time from Cardiac Magnetic Resonance Imaging Data*

Kateřina Škardová

5th year of PGS, email: `katerina.skradova@fjfi.cvut.cz`
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. In this contribution, a method for estimating tissue parameters using cardiac magnetic resonance imaging (MRI) and a biophysical model by combining neural network (NN) and numerical optimization (NO) is presented. The method is used on the problem of T_1 relaxation time estimation from the Modified Look-Locker Inversion recovery (MOLLI) image series. The of MRI used in this work is based on Bloch equation.

MOLLI data were acquired for eight phantoms (with varying T_1 relaxation time) on 1.5 and 3T MRI systems; twelve patients on 1.5T system and three patients on 3T system (native and after administering MRI contrast agent reducing T_1). In the phantom study, additionally, inversion recovery turbo spin echo (IR-TSE) data were acquired and provided a pseudo-ground truth of the T_1 relaxation time (T_1^{pGT}).

T_1 from MOLLI images was obtained by the scanner ($T_1^{scanner}$) and by the proposed method, which consists of two stages: the first-stage estimate T_1^{NN} is given by the NN trained on synthetic data generated by the Bloch simulator, the final estimate $T_1^{NN,NO}$ by numerical optimization (NO). After the validation on the phantom data, the T_1 maps were created from routine cardiac MRI MOLLI data by the proposed two-stage method and compared with the T_1 map provided directly by the scanner.

The proposed two-stage method provided results comparable to the scanner in phantoms study on 1.5T and results closer to pseudo-ground truth in 7 out of 8 phantoms on 3T. The NO stage improved the accuracy and decreased the variation of T_1 obtained by the NN stage.

For the *in vivo* data, the reference values computed based on IR-TSE were not avaliable. The pre-contrast $T_1^{NN,NO}$ in blood and myocardium was higher than $T_1^{scanner}$ in most subjects measured on 1.5T and 3T system. The post-contrast $T_1^{NN,NO}$ was higher than $T_1^{scanner}$ in blood for subject measured on both systems. In the case of post-contrast myocardium, $T_1^{scanner}$ was higher than $T_1^{NN,NO}$ in most subjects measured, which is in line with the phantom experiment and expected due to known underestimation of T_1 from MOLLI data.

*This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic under the OP RDE grants number CZ2.11/0/0/16_019/0000778 “Centre for Advanced Applied Sciences”, CZ2.11/0/0/16_019/0000765 “Research Center for Informatics”, by Ministry of Health of the Czech Republic project No. NV19-08-00071 and by Institutional Support MHCZ-DRO (“Institute for Clinical and Experimental Medicine – IKEM, IN 00023001”). This work was also supported by the Inria France-UT Southwestern Medical Center Dallas Associated Team TOFMOD.

NO initialized by the NN, which can be trained using simulated data, has the potential to increase the efficiency and robustness of tissue parameter estimation from image data.

Keywords: T_1 relaxation time, Magnetic Resonance Imaging (MRI), Modified Look-Locker Inversion recovery (MOLLI), Parameter estimation, Bloch simulator

Abstrakt. V tomto příspěvku je představena metoda pro odhad parametrů tkáně na základě dat z magnetické rezonance (MRI) a biofyzikálního modelu MRI, za použití kombinace neuronové sítě (NN) a numerické optimalizace (NO). Metoda je aplikována na problém odhadu T_1 relaxačního času z MOLLI (Modified Look-Locker Inversion Recovery) obrazových dat. Použitý model MRI je založený na Blochových rovnicích.

MOLLI data byla naměřena pro osm fantomů (s různým T_1 relaxačním časem) na přístrojích o síle pole 1.5 a 3T; dvanáct pacientů na přístroji o síle pole 1.5T a tři pacienti na přístroji o síle pole 3T (nativně a po podání kontrastní látky snižující T_1). Ve fantomové studii byla navíc naměřena data pomocí Inversion Recovery Turbo Spin Echo (IR-TSE) sekvence. Data naměřená pomocí IR-TSE sekvence byla použita pro výpočet referenční hodnoty T_1 relaxačního času (T_1^{pGT}).

Při použití navržené metody je T_1 relaxační čas z MOLLI dat získán ve dvou krocích: první odhad T_1^{NN} je získán pomocí NN trénované na syntetických datech generovaných Blochovým simulátorem, finální odhad $T_1^{NN,NO}$ je získán pomocí numerické optimalizace (NO). Při validaci na fantomech byly výsledky navržené dvou krokové metody porovnány s hodnotami odhadnutými na základě MOLLI dat přímo MR skenerem ($T_1^{scanner}$) a s referenčními hodnotami určenými na základně IR-TSE sekvence.

Navržená metoda poskytla výsledky srovnatelné se skenerem ve studii na fantomech měřených na stoji o síle pole 1.5T a výsledky bližší referenční hodnotě pro 7 z 8 fantomů v případě mření na stroji o síle pole 3T. Srovnání ukázalo že druhý krok (NO) zlepšil přesnost a snížil rozptyl T_1 získaných v prvním kroku navržené metody (NN).

U *in vivo* měření nebyla k dispozici referenční hodnota T_1. Předkontrastní hodnota $T_1^{NN,NO}$ v krvi a myokardu byla vyšší než $T_1^{scanner}$ u většiny subjektů měřených na obou MR strojích. Postkontrastní hodnota $T_1^{NN,NO}$ v krvi byla vyšší než $T_1^{scanner}$ u všech subjektů měřených na obou strojích. V případě postkontrastní hodnoty v myokardu byla $T_1^{scanner}$ u většiny měřených subjektů vyšší než $T_1^{NN,NO}$, což je v souladu s fantomovým experimentem.

Inicializace druhého kroku numerické optimalizace pomocí odhadu získaného neuronovou sítí, kterou lze trénovat pomocí generovaných dat, má potenciál zvýšit účinnost a robustnost odhadu parametrů tkáně z obrazových dat.

Klíčová slova: T_1 relaxační čas, Magnetická rezonance, Modified Look-Locker Inversion recovery (MOLLI), Odhad parametrů, Blochův simulátor

Full paper: Kateřina Škardová, Radek Galabov, Kateřina Fricková, Tomáš Pevný, Jaroslav Tintěra, Tomáš Oberhuber, Radomír Chabiniok *Combining machine learning and mathematical modeling in estimation of T1 relaxation time from cardiac magnetic resonance imaging data.* The paper is currently under review in Magnetic Resonance Imaging.

Numerická optimalizace Dirichletovy okrajové podmínky pro problém fázového pole s aplikací na růst krystalu [*]

Aleš Wodecki

3. 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 pure substance solidification of a supercooled melt using the phase field model. This model is composed of the heat equation and the phase field equation coupled with suitable initial and boundary conditions, one of which is controlled. As opposed to the distributed control of parabolic PDE's, very few contributions currently exist pertaining to the Dirichlet boundary condition control for parabolic PDE's. This motivates our interest in the Dirichlet boundary condition control for the phase field model describing the solidification of a pure substance from a supercooled melt. In particular, our aim is to control the time evolution of the temperature field on the boundary of the computational domain in order to achieve the prescribed shape of the crystal at the given time. To obtain efficient means of computing the gradient of the cost functional, we derive the adjoint problem formally. The gradient is then used to perform gradient descent. The viability of the proposed optimization method is supported by several numerical experiments performed in one and two spatial dimensions. Among other things, these experiments show that a linear reaction term in the phase field equation proves to be insufficient in certain scenarios and so an alternative reaction term is considered to improve the models behavior.

Keywords: phase field, anisotropic crystal growth, optimization, Dirichlet boundary condition

Abstrakt. Za užití modelu fázového pole simulujeme solidifikaci čisté směsi. Použitý model se skládá z rovnice vedení tepla a rovnice fázového pole a je doplněn vhodnými okrajovými a počátečními podmínkami, jedna z nich je řízena. Na rozdíl od distribuovaného řízení parabolických rovnic, jež je hojně studováno, jen malé množství příspěvků se věnuje řízení pomocí Dirichletovy okrajové podmínky. Toto motivuje náš zájem o řízení rovnic fázového pole, popisujících tuhnutí čisté směsi, s pomocí Dirichletovy okrajové podmínky. Cílem tohoto optimálního řízení je nalezení Dirichletovy okrajové podmínky pro rovnici vedení tepla, která vede k předepsanému tvaru fázového pole v konečném (předem daném) čase. Abychom zajistili efektivní výpočet gradientu využijeme formálního odvození adjugovaných rovnic. Gradient je potom využit abychom provedli gradientní sestup. Validita této techniky je ověřena s pomocí numerických experimentů

[*] Tato práce byla podpořena grantem grantové agentury ČVUT v Prague, grant No. SGS17/194/OHK4/3T/14...

v jedné a dvou dimenzích. Mimo jiné je v těchto experimentech studován vliv různých reakčních členů. Při těchto experimentech se ukazuje, že existují řízení, při nichž je potřeba užít pokročilejších reakčních členů, aby byla zajištěna realističnost výsledného řízení.

Klíčová slova: rovnice fázového pole, anisotropický růst krystalů, optimalizace, Dirichletova okrajová podmínka

Plná verze: A. Wodecki, P. Strachota, T. Oberhuber, K. Škardová, M. Balázsová. *Numerical Optimization of the Dirichlet Boundary Condition in the Phase Field Model with an Application to Pure Substance Solidification.* https://arxiv.org/abs/2208.13910

Literatura

[1] R. Eymard. *Finite volume methods, Handbook of Numerical Analysis, vol. 7.* Elsevier, 2000, pp. 715–1022.

[2] 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

[3] P. Strachota. *Focusing the latent heat release in 3D phase field simulations ofdendritic crystal growth.* Modelling Simul. Mater. Sci. Eng. 29 065009 (2021). ISSN: 0965-0393. https://doi.org/10.1088/1361-651X/ac0f55.

Non-equilibrium Strain Induces Hysteresis and Anisotropy in the Quasi-Static and Dynamic Elastic Behavior of Sandstones: Theory and Experiments

Radovan Zeman

2nd year of PGS, email: zemanra5@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Zdeněk Převorovský, Department of Impact and Waves in Solids
Institute of Thermomechanics, CAS

Jan Kober, Department of Impact and Waves in Solids
Institute of Thermomechanics, CAS

Abstract. Materials with grain contacts or partially closed cracks exhibit anomalous elastic behavior: hysteresis in quasi-static experiments and slow dynamics in fast dynamic ones. When a slowly varying strain is applied, the wave velocity increases with incresing loading and hysteresis is observed. In the dynamic range, a forcing which varies rapidly in time is applied and non-equilibrium phenomena are observed. The system recovers its initial state when the conditioning is removed. Albeit the behavior in the two cases (which correspond to very different strain ranges) looks different, it should stem from the same physics and thus could be modelled by the same equation of state. Here, we propose a modification of the standard acoustoelastic theory, introducing the concept of conditioning induced non-equilibrium strain, which is defined correctly for each experiment and results in hysteris and slow dynamics. The resulting model allows to predict the behavior in both quasi-static and dynamic ranges, including velocity anisotropy induced by nonlinearity.

Keywords: nonlinear elasticity, consolidated granular materials, locked-in stress, acoustoelastic testing

Abstrakt. Materiály obsahující kontakty zrn nebo částečně uzavřené defekty vykazují anomální elastické chování: hysterezi v kvazistatických experimentech a slow dynamics v dynamických experimentech. Při aplikaci pomalu se měnící deformace se rychlost vlnění zvyšuje s rostoucím zatížením a je pozorována hystereze. V dynamickém testování se aplikuje síla, která se rychle mění v čase, a jsou pozorovány nerovnovážné jevy. Po odstranění conditioningu se systém vrací do původního stavu. Přestože chování v obou případech – které odpovídají velmi rozdílným rozsahům deformace – vypadá odlišně, mělo by vycházet ze stejného fyzikálního principu, a proto by mělo být modelováno stejnou stavovou rovnicí. Zde navrhujeme modifikaci standardní akustoelastické teorie, která zavádí koncept nerovnovážné deformace vyvolané conditioningem, která je správně definována pro každý experiment a způsobuje hysterezi a slow dynamics. Výsledný model umožňuje předpovídat chování v kvazistatickém i dynamickém rozsahu, včetně anizotropie v rychlostech vln vyvolané nelinearitou.

Klíčová slova: nelineární elasticita, konsolidované granulované materiály, locked-in napětí, akustoelastické testování

Full paper: J. Kober, M. Scalerandi, R. Zeman, *Non-equilibrium strain induces hysteresis and anisotropy in the quasi-static and dynamic elastic behavior of sandstones: Theory and Experiments.* Submitted to Applied Physics Letters.