DOKTORANDSKÉ DNY 2023

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

10. a 24. listopadu 2023

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2023
sborník workshopu doktorandů FJFI oboru Matematické inženýrství

P. Ambrož, Z. Masáková (editoři)
Kontakt petr.ambroz@fjfi.cvut.cz / 770 127 206

Vydalo České vysoké učení technické v Praze
Zpracovala Fakulta jaderná a fyzikálně inženýrská

Počet stran 162, Vydání 1.

Seznam příspěvků

Předmluva

Doktorandské dny jsou již tradičním setkáním studentů doktorského studia na Fakultě jaderné a fyzikálně inženýrské ČVUT v Praze. Zúčastňují se ho doktorandi studijních programů Matematické inženýrství a Aplikovaná informatika, na jejichž zajištění se podílejí katedry matematiky, fyziky a softwarového inženýrství. Studenti prezentují výsledky své vědecké práce, jejichž tematika pokrývá všechny oblasti aplikované matematiky.

Letošní ročník je již osmnáctým vydáním workshopu, koná se ve dnech 10. a 24. listopadu 2023 v prostorách FJFI.

Tento sborník přináší jak plné texty studentských příspěvků, tak i abstrakty s odkazy na články otištěné 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 33/23/F4.

Editoři

Manifold Analysis in the Coreference Resolution Problem*

Vladislav Belov
belovvla@cvut.cz

study programme: Mathematical Engineering
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 can worsen the output quality and are required to be revised. In most exercises with real datasets, points are distributed along a highly nonlinear manifold whose structure is unknown. In this paper, we aim to analyze the Coreference Resolution (CR) problem from the perspective of Manifold Learning. We investigate the influence of dimensionality reduction on the neural CR framework performance and empirically verify the hypothesis that one can affect the efficiency in both positive and negative directions quality-metric-wise. Furthermore, we identify coreferent clusters in the OntoNotes 5 dataset and examine their evolution during the framework training process, showcasing that the current CR architectures do not account for separating such clusters. With that, we open a slot for further research.

Keywords: coreference resolution, dimensionality reduction, manifold learning, noise reduction

Abstrakt. Vysoký počet dimenzí může způsobit různé problémy v reálných aplikacích. Některé rozměry mohou zhoršit kvalitu výstupu a je nutné je revidovat. Kromě toho, při práci se skutečnými daty se často setkáváme s případy, když jsou distribuována podél nějaké nelineární variety, jejíž struktura není známa. V tomto příspěvku se zaměřujeme na analýzu problému rozlišení koreferencí (angl. Coreference Resolution - CR) z pohledu varietního učení. Zkoumáme vliv redukce dimenzionality na výkon neuronové architektury CR a empiricky ověřujeme hypotézu, že lze ovlivňovat efektivitu v pozitivním i negativním smyslech z hlediska kvality modelu. Dále identifikujeme koreferenční clustery v datové sadě OntoNotes 5 a zkoumáme jejich vývoj během učení, přičemž ukazujeme, že současné architektury CR nepočítají s prostorovým oddělením takových clusterů, a otevíráme prostor pro další výzkum.

Klíčová slova: redukce dimenzionality, redukce šumu, rozlišení koreferencí, varietní učení

1 Introduction

Modern Natural Language Processing (NLP) approaches can achieve significant results in standard textual analysis tasks. The list of tasks includes but is not limited to such fields as text classification, e.g., determining the general topic of the news article [1] or determining the text author's attitude towards the topic [22]; sequence tagging, e.g.,

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS23/187/OHK4/3T/14.

named entity recognition (NER) [29, 37, 35, 19] and part-of-speech tagging [7]; and text generation [13]. For some applications, it is essential to combine these tasks to achieve more comprehensive results. For instance, in the general case of the sentiment analysis task, one aims to classify whether the author of the piece of text refers to the topic in a negative or positive sense. However, to obtain an adequate understanding of why their attitude is inferred to be equal to a particular value, it is essential to discern contextual dependencies within the piece, especially in cases when the range of output values goes beyond "polarity" (positive/neutral/negative) and matures into a broader spectrum of values like doubt, contempt, or enjoyment. The NLP research community not only actively develops models for improved natural language understanding, i.e., representing language in a vector space, [26, 9, 36] but also proposes different fine-tuning approaches for these models [18, 14, 27, 8].

In this work, we aim to perform research on constructing better contextual dependencies in the form of improved Coreference Resolution (CR). The CR task is rather complex to solve. Starting in 2017 [17] to this date [10], research activity in the field has significantly advanced thanks to neural architectures. Even though improvements on the standard benchmark [25, 33] have been relatively slight, the community pushed the target F1 score to 81.0%. Nonetheless, this value is still imperfect, as it is insufficient for real-world applications. Hence, the range of possibilities for new solutions is far enough from exhausting and landing at the saturation point of the research.

Coreference Resolution combines detection and linking various mentions of entities within the text: linking noun phrases with their counterparts and pronouns, anaphora disambiguation, linking words with their pro-forms, etc. Hence, CR-solving models significantly impact the quality of the text-mining algorithms. A good use case where coreference resolution can be applied is categorizing entities and their pronouns to provide one with a broader spectrum of information for future decision-making. Based on the extracted data, it is possible to unify all knowledge in the form of a Knowledge Graph (KG) [31], which can be further utilized for linking concepts represented by textual spans. Dependencies and connections between the entities can enrich the feature space with highly discriminative samples for other tasks. For example, we can assume having the following two consecutive sentences: "John Smith and Amanda Brown are employees of company XYZ. Amanda's colleague was accused of drunk driving." Based on these sentences, one could desire to classify if any of the entities from the text can be charged with a misdemeanor. For a human reader, it is evident that span "Amanda's colleague" refers to John. However, for a machine, that is a challenging task. Therefore, proper identification of entity clusters like {"John Smith", "Amanda's colleague"}, {"Amanda Brown"}, and {"XYZ"} would significantly improve the machine's understanding of the piece of text.

Existing CR models do not explore the topic of which dimensions to choose for modeling; only raw high-dimensional context-dependent vector representations are passed to CR models. We aim to investigate this aspect to analyze and enhance model learning capabilities. Additionally, the optimization of vector embeddings outputted by complex Transformer-based models may allow us only to use their pre-trained weights and focus on training a lighter architecture. In [5], we explored existing nonlinear dimensionality reduction techniques and investigated the ability to measure the quality of Manifold Learning methods. Most of the time, some dimensions are redundant and carry no mean-

ingful information. This noise unsatisfactorily affects the output of machine learning algorithms due to overfitting. In addition, if data points are distributed along an unknown manifold with a nonlinear structure whose actual dimension is smaller than the original space, then even a simple comparison of pairwise Euclidean distances will be insufficient. In such cases, one must approximate geodesics on the manifold to obtain reliable results.

This work is structured straightforwardly: Sec. 2 comprises our knowledge of state of the art, where we review the Coreference-Resolution-related literature and the concept of nonlinear dimensionality reduction; Sec. 3 addresses the topic of neural Coreference Resolution in greater detail; in Sec. 4 and Sec. 5, we analyze the embedding spaces of the selected Coreference Resolution framework and discuss experimental results, respectively. Finally, Sec. 6 concisely summarizes this paper.

2 Related Work

Coreference Resolution

Modern Coreference Resolution algorithms are combinations of sophisticated vector embeddings representing context and deep neural network superstructures performing the coreference resolution. The first end-to-end neural coreference resolution model was introduced in [17]. Its crucial difference from its predecessors was that it did not require preprocessing in the form of syntactic parsing or rule-based mention detection since the model can learn to mention dependencies on its own to a forerunner-outperforming extent. The main idea of the model is to learn to score pairs of textual spans in such a way that takes into account, firstly, if these spans are entity mentions and, secondly, whether the pair is of type antecedent-descendant in terms of coreference. The NLU model of choice provides span representations. The goal is to assign to each span an antecedent span. [14] belongs to the state-of-the-art approaches that utilize the same structure on top of SpanBERT. One of the crucial drawbacks of the scoring approach is the choice of spans: sizes of relevant spans can be different, so a constant width of the window may not always be the right choice; spans can either overlap or be disjoint; if they overlap, the value of the overlap also becomes a hyperparameter. In addition, the number of scoring procedures is quadratic in complexity: each span has to be scored against all of its counterparts. If the length of the document is significant, the memory needed to store all entity mentions may become an issue (in [34], authors propose an incremental structure for the CR model, which requires a lot less memory for the price of a slight decrease in performance). While previous models can achieve decent results, their memory footprint is still noticeable. The authors of [15] bypassed the need to create span representations, relying on a combination of bilinear functions applied on endpoint token representations. In addition, the new model is built on top of a Longformer encoder capable of processing long documents. Furthermore, the World-Level coreference resolution model based on RoBERTa [18] appeared with its novel approach, bypassing the necessity of span representation and computing coreference scores between individual tokens instead. The model from [10] went a step further and completely omitted the step of span representations, evaluating coreference scores on words. A different outlook on the CR problem

is provided by Google's sequence-to-sequence approach [6]. Their model transforms the task from modeling the coreference scores to generating sequences with additional special tokens from original texts. These special tokens are designed to represent coreference clusters.

Manifold Learning

One of the earliest known Manifold Learning techniques, built on top of the classical Multidimensional Scaling, is Isomap [30]. It utilizes shortest paths on k-Nearest-Neighbors graphs of data to preserve geodesics. At the same time as that of Isomap, Locally Linear Embedding (LLE) [28] was introduced with the proposal to express each point as a linear combination of its neighbors. The objective is to preserve these linear combinations in the latent space. Other noteworthy methods are, for example, the Laplacian Eigenmaps (Spectral Embedding) [3, 4] with its point proximity preservation using the Laplacian of k-Nearest-Neighbors graphs and the Hessian Locally Linear Embedding (HLLE) [11] which is considered to be not only an LLE modification based on second derivations but also a projection technique mathematically closely related to the Laplacian Eigenmaps. All of the abovementioned techniques can be considered "pure" Manifold Learning, as they all aim to approximate geometric properties of original data in the latent space. For instance, a completely different approach is introduced in the article on the t-Distributed Stochastic Neighbor Embedding (t-SNE) [20], a method aimed at preserving the probability distribution of point neighborhoods. Another prominent technique is the Uniform Manifold Approximation and Projection (UMAP) [21], built upon vast mathematical fields such as topology and category theory. TriMap [2] is a technique based on triplet constraints. The method is aimed at scoring designed to reflect the relative position of clusters instead of individual points. Last but not least, we mention the Manifold Learning method that is shown to be superior to its predecessors both in performance and computation time - PaCMAP [32], an algorithm born from a comprehensive analysis of other methods. The authors designed a particular loss function that is minimized in multiple stages and accounts for near-pairs, mid-near pairs, and further pairs in terms of points and k-neighborhoods.

3 Neural Coreference Resolution Models

Starting with basic notation, let D refer to a textual document. Each document contains T words and, therefore, $N = \frac{T(T+1)}{2}$ possible spans. In addition, documents may contain metadata such as speaker information or genre.

The main focus of a coreference resolution model is to assign to each span i an antecedent span $y_i \in Y(i) = \{\varepsilon, 1, 2, \ldots, i-1\}$. The definition of $Y(i)$ clearly contains all spans to the left of i.[1] Span ε is a special antecedent reserved for empty coreference relations.

The key to teaching a model to predict antecedents is to learn a conditional probability that produces the correct clustering of antecedents:

[1] We consider left-to-right languages.

$$P\left(y_1, y_2, \ldots, y_N|D\right) = \prod_{i=1}^{N} P\left(y_i|D\right) = \prod_{i=1}^{N} \frac{e^{\,s(i,y_i)}}{\sum_{y' \in Y(i)} e^{\,s(i,y')}}, \tag{1}$$

Where $s(i, \cdot)$ is the pairwise coreference score for span i and its potential antecedent from $Y(i)$. Since ε denotes empty coreference, $s(i, \varepsilon) = 0$, for all i. In this text, all relations will be assumed to be connected to a specific document - for this reason, the conditionality concerning D will be omitted. It is important to note that major differences between various models stem from how scoring function s is constructed.

In the learning phase, as the antecedents are latent, the minimization of the marginal log-likelihood of only correct antecedents implied by the gold clustering is assumed:

$$L_{\,\mathrm{COREF}} = -\log \prod_{i=1}^{N} \sum_{y' \in Y(i) \cap \mathrm{GOLD}(i)} P(y'). \tag{2}$$

In (2), GOLD(i) refers to the set of spans in the gold cluster mentioning span i. If i is not found in any of the gold clusters or its true antecedents have been pruned during the scoring, GOLD(i) = $\{\varepsilon\}$. With such a learning objective, only gold mentions undergo positive updates when scores of non-gold antecedents are pushed lower.

The so-called end-to-end (E2E) model described in [17] computes the coreference score as follows:

$$s(i,j) = \begin{cases} s_{\mathrm{m}}(i) + s_{\mathrm{m}}(j) + s_{\mathrm{a}}(i,j), & \text{if } j \neq \varepsilon, \\ 0, & \text{if } j = \varepsilon. \end{cases} \tag{3}$$

Where $s_{\mathrm{m}}(\cdot)$ is the score component evaluating whether a span is an entity mention and $s_{\mathrm{a}}(i,j)$ is the pairwise score assessing whether span j is an antecedent of span i.

The main difference between the start-to-end (S2E) [15] and the E2E [17] models is based on the fact that the S2E model avoids computing span representations. Instead, the mention s_{m} and antecedent s_{a} scores are proposed to be computed utilizing the series of linear transformations over span boundaries, i.e., start and end tokens.

Word-Level (WL) Coreference Resolution

The number of spans in the document is $O(T^2)$; therefore, the number of potential coreference links between spans is $O(T^4)$. While introducing enhancements such as the mention pruning to the existing framework, other neural models, such as E2E and S2E, still need to address the complexity issue, as they are fully span-dependent. The WL model [10] proposes to evaluate coreference links between individual words, reconstructing spans afterward. This step facilitates the reduction of the model complexity to $O(T^2)$. Such reduction enables us to consider all potential coreference links without pruning out any of the candidates.

The WL model coreference scores $s(i,j)$ consist of only two factors: the coarse and the fine antecedent scorers. Therefore, it is computed as the following straightforward sum:

$$s(i,j) = s_{\mathrm{c}}(i,j) + s_{\mathrm{a}}(i,j). \tag{4}$$

Consequently, the span extraction model is applied to the tokens found to be coreferent. The span model reconstructs spans by predicting their most probable start and end tokens.

Furthermore, the WL model introduces an additional term as a regularization factor to the main loss function to encourage the model to output higher coreference scores for all correct coreferent mentions:

$$L = L_{\text{COREF}} + \alpha L_{\text{BCE}}. \tag{5}$$

The value of α is originally put to 0.5 to prioritize the primary coreference loss.

4 Manifold Analysis of CR Frameworks

This section briefly describes our approach to the manifold-learning-based analysis of the CR frameworks. In this paper, we aimed primarily at the WL CR model as a target for the analysis due to the availability of the code base and ease of deployment on the RCI

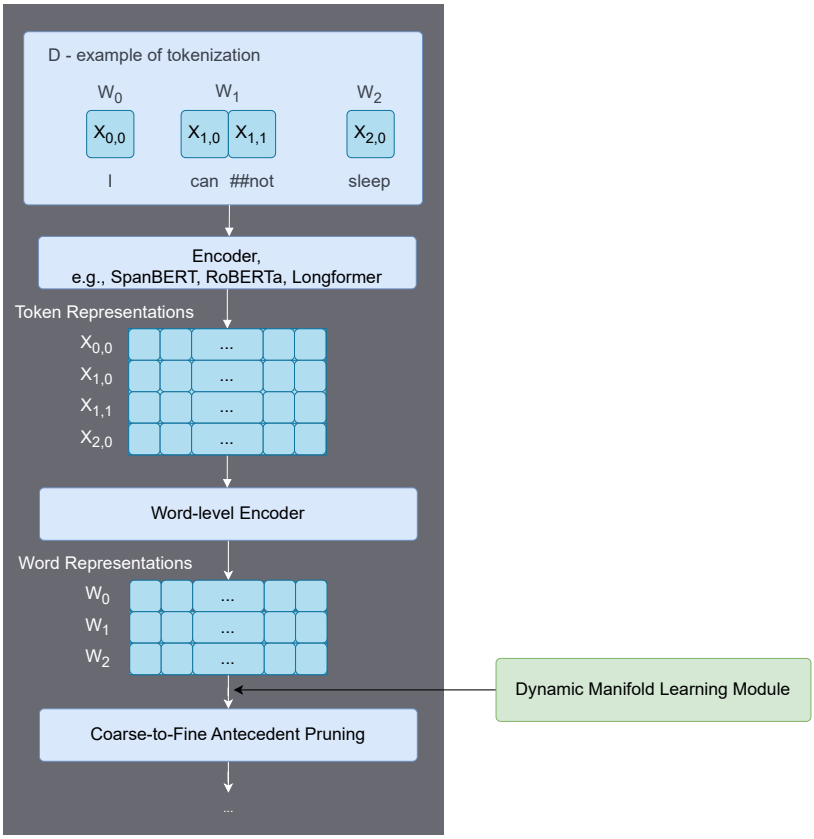


Figure 1: Diagram depicting the design of the experiment aiming to influence the outcome of the WL Coreference Resolution model through dimensionality reduction applied on word embeddings. The green box depicts the placement of the influencing module introduced by us.

infrastructure.[2]

The first analysis we carried out was aimed at the ability to influence the outcome of the model predicting capabilities through applying dimensionality reduction to the layers of word-level embeddings. Fig. 1 partially depicts the design of the experiment. The critical assumption is that the initial embedding space containing the mention clusters is noisy and contains redundant dimensions. Dimensionality reduction can either denoise the space, making it easier for the model to represent antecedent probabilities, or reduce the space excessively, causing the model to lose information about antecedents. Since each Manifold Learning method involves minimizing a kind of loss function, the WL model is also required to consider this loss. Therefore, we modified (5) as follows:

$$L = L_{\text{COREF}} + \alpha L_{\text{BCE}} + \beta L_{\text{MANIFOLD}} \tag{6}$$

Where β represents the aggressivity of the manifold-learning-based loss.[3] For each run, we measured the LEA F1 Score [23] on the test set of the OntoNotes 5 corpus [25, 33] and compared it to the original model performance.

The second analysis once again concerns the word embeddings. This time, from a different perspective: since the WL model is computing coreference scores for headwords, one can straightforwardly analyze the relationship between the headword embeddings and their connection to particular coreference clusters within the vector space. After each epoch, we extract word embeddings for each document, reduce the dimensionality of the input space,[4] and map it onto a two-dimensional plane via PaCMAP technique. Additionally, we measure the average in-cluster distance within the original space. We hypothesize that the input word-level embedding space is being transformed during training in favor of closer distance for coreferent mentions.

5 Experimental Results

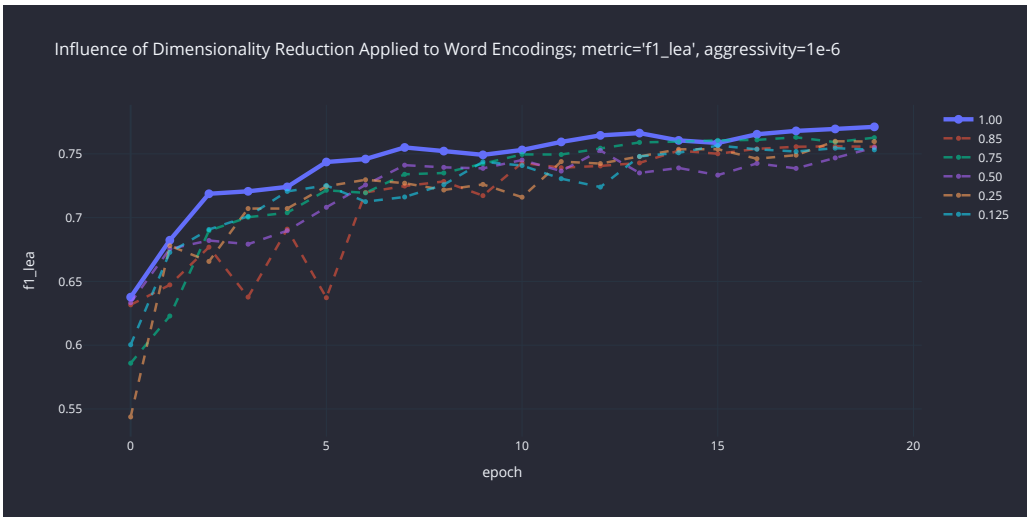
For the first experiment involving influencing the outputs of the CR model with dimensionality reduction (for simplicity, we utilize PCA-like reconstruction loss) applied to word embeddings, our findings are displayed in Fig. 2. In Fig. 2a, one can see that for $\beta = 10^{-6}$ dimensionality reduction results in loss of predicting capabilities, as all variants underperform in comparison with the original model. However, a different trend may be observed in Fig. 2b. Aggressivity $\beta = 10^{-7}$ results in scores close to original values, in some epochs even over-performing its counterpart. The causing factor of such performance is the relative increase in precision (see Fig. 3). One can speculate that dimensionality reduction reduces the noise in this case and introduces a new embedding space, which is more favorable for higher precision values.

The results for the second analysis are depicted in Fig. 4, where we selected two

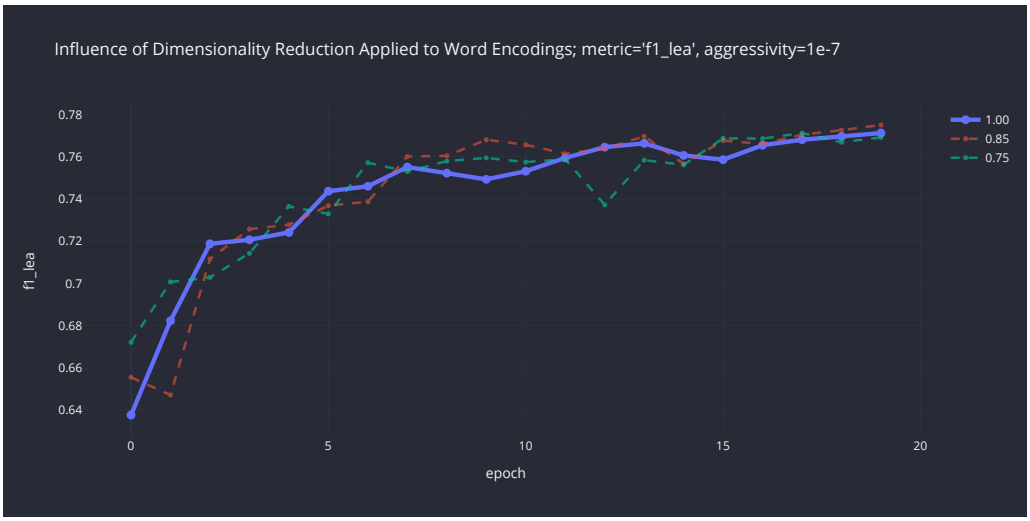
[2]The access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics" is also gratefully acknowledged.

[3]In our experiments, we found $\beta \in [10^{-7}, 10^{-6}]$ to be sufficient for scaling component magnitudes evenly.

[4]The output dimensionality of the word-level encoder is set to 1024.



(a) Aggressivity $\beta = 10^{-6}$



(b) Aggressivity $\beta = 10^{-7}$

Figure 2: Evolution of LEA F1 Scores on the test set of OntoNotes 5 with different settings of L_{MANIFOLD} aggressivity. On the right-hand side, the legend depicts the percentage of dimensions remaining after the dimensionality reduction: 1.0 is for the original model with no dimensionality reduction involved, 0.85 stands for the 85% preservation of the dimension count, etc.

documents and displayed two-dimensional PaCMAP mapping results.[5] In the following Fig. 5, we display the epoch-wise evolution of average pairwise in-cluster distances. Both figures indicate that the space does not reflect coreference clusters initially since the positioning of coreferent headwords is spontaneous. However, with time, the model starts to adopt the behavior enforced by the coreference loss, pushing coreferent headwords closer to each other. Even though we observe such an act, one can also notice that the average in-cluster distance stabilizes quickly after a few epochs, and no significant changes

[5]Similar behavior is observed for other documents as well.

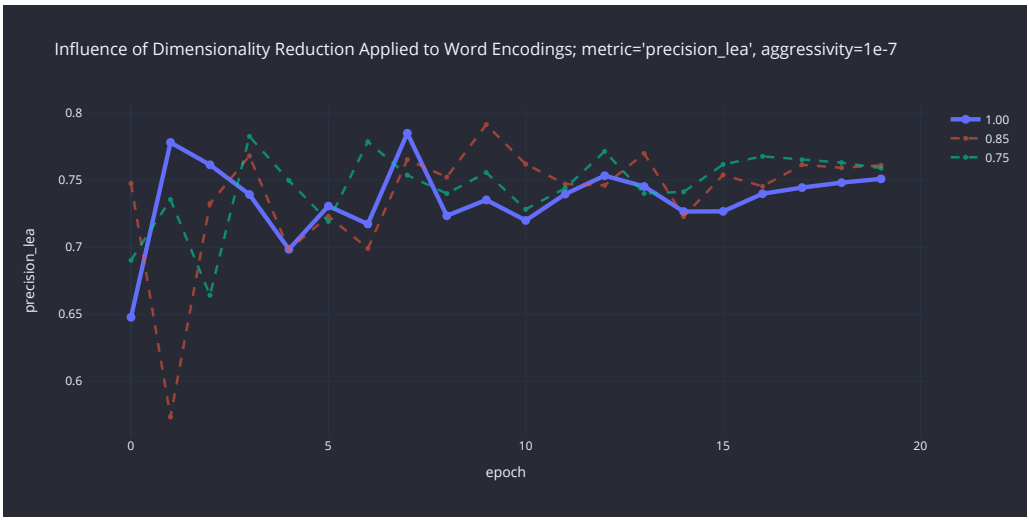


Figure 3: Evolution of LEA Precision Scores on the test set of OntoNotes 5 for the settings of Fig. 2b.

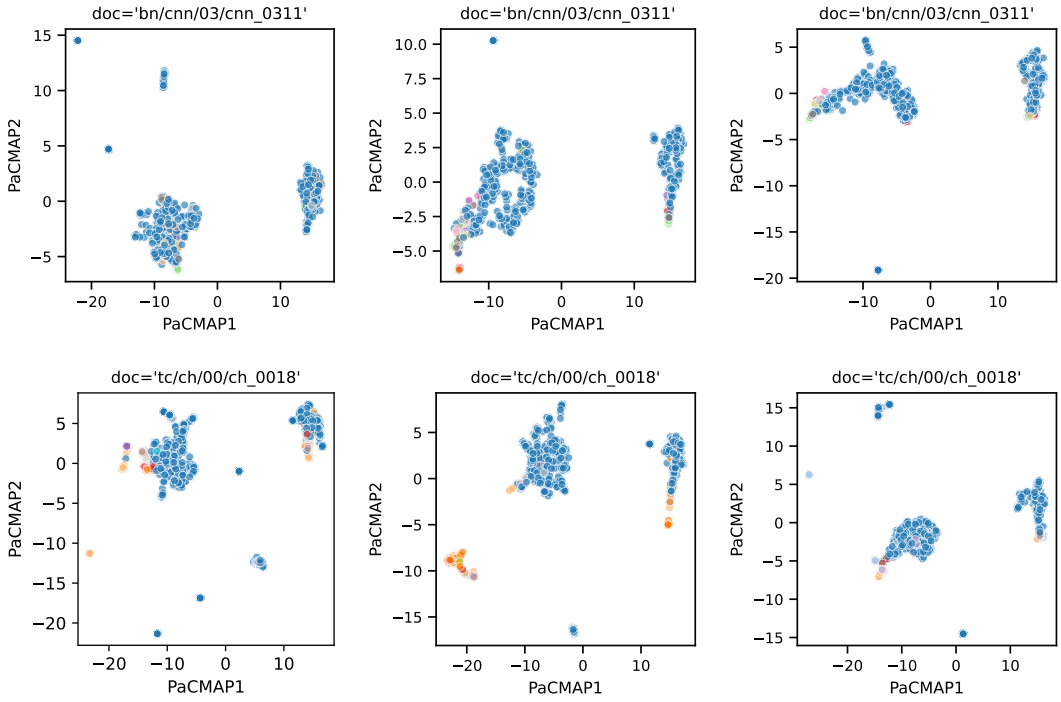
occur, while the model learning curve also achieves saturation (recall the lines for the LEA F1 Score from Fig. 2). Hence, one can speculate that by improving the embeddings further, the model can perform even better, separating the coreference clusters from each other, enhancing mention scoring and the overall model performance statistics.

6 Conclusion

This paper reviewed the progress in neural Coreference Resolution and Manifold Learning fields in Sec. 2. We described neural-network-based approaches in Sec. 3: the foundation end-to-end architecture and the superstructures built upon it. We focus on the Word-level CR framework to perform the manifold-learning-based analysis. In Sec. 4 and 5, we investigate the influence of dimensionality reduction on LEA scores and the behavior of the original word embedding space with respect to coreference clusters. We conclude that dimensionality reduction can influence the performance of the models in both positive and negative ways. Additionally, through nonlinear dimensionality reduction, we could examine the original high-dimensional embeddings and deduce that the learning process halts significant changes concerning coreferent mentions, potentially affecting the early saturation of LEA F1 behavior. The approaches proposed in this paper introduce the necessity and reason to pay more attention to underlying semantic representations from NLU models and to research further potentials of improving the architecture, allowing more efficient separation of coreferent mentions.

In the future, we aim to analyze the impact of more complex Manifold Learning structures on the (sub-)word level. Additionally, the recent surge in the NLP community around Generative AI capabilities opens doors for new approaches and improved contextualized vectorization. For instance, it has been shown that embeddings from hypothetical descriptions (AI-generated descriptions that do not have to be factually correct) lead to better performance when semantic inference is of interest [12]. Moreover,

InstructGPT [24] performed well as a zero-shot coreference resolution system [16]. We aim to deeply analyze the influence of such systems on the coreference resolution inference and underlying spatial transformations.



(a) Epoch 0 (b) Epoch 4 (c) Epoch 19

Figure 4: PaCMAP visualizations of coreference clusters over different training epochs for two OntoNotes 5 documents. The blue color depicts non-coreferent document words. Other colors represent individual coreference clusters.

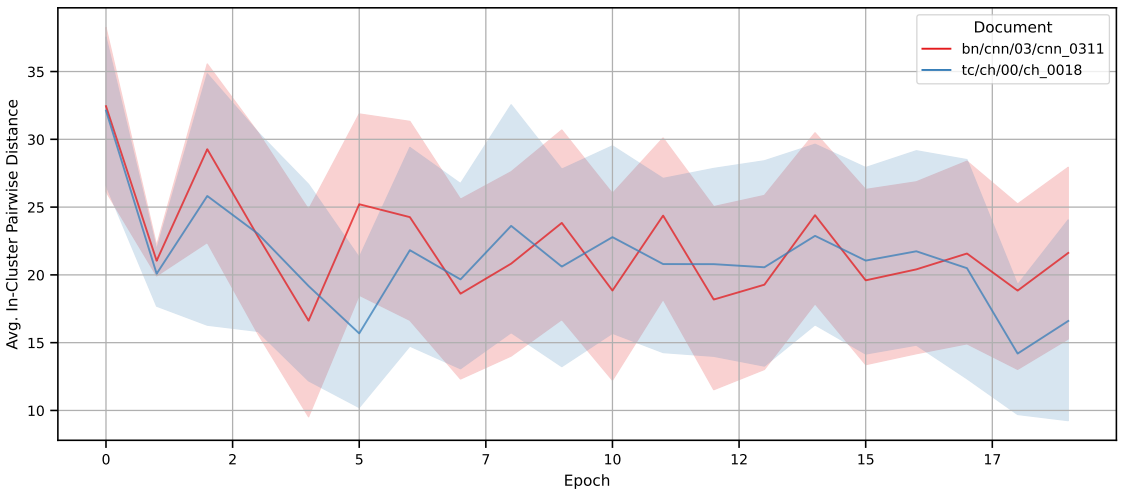


Figure 5: In-cluster distance evolution over different training epochs for two OntoNotes 5 documents from Fig.4.

References

[1] B. Altinel and M. C. Ganiz. *Semantic text classification: A survey of past and recent advances.* Information Processing & Management **54** (2018), 1129–1153.

[2] E. Amid and M. K. Warmuth. *TriMap: Large-scale Dimensionality Reduction Using Triplets.* CoRR **abs/1910.00204** (2019).

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In 'NIPS', volume 14, 585–591, (2001).

[4] M. Belkin and P. Niyogi. *Laplacian eigenmaps for dimensionality reduction and data representation.* Neural computation **15** (2003), 1373–1396.

[5] 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.

[6] B. Bohnet, C. Alberti, and M. Collins. Coreference resolution through a seq2seq transition-based system, (2022).

[7] B. Bohnet, R. McDonald, G. Simões, D. Andor, E. Pitler, and J. Maynez. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', 2642–2652, Melbourne, Australia, (July 2018). Association for Computational Linguistics.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, (2020).

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, (2019).

[10] V. Dobrovolskii. *Word-Level Coreference Resolution.* CoRR **abs/2109.04127** (2021).

[11] D. L. Donoho and C. Grimes. *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data.* Proceedings of the National Academy of Sciences **100** (2003), 5591–5596.

[12] L. Gao, X. Ma, J. Lin, and J. Callan. Precise Zero-Shot Dense Retrieval without Relevance Labels, (2022).

[13] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang. Long text generation via adversarial training with leaked information, (2017).

[14] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. *Spanbert: Improving pre-training by representing and predicting spans.* CoRR **abs/1907.10529** (2019).

[15] Y. Kirstain, O. Ram, and O. Levy. *Coreference Resolution without Span Representations.* CoRR **abs/2101.00434** (2021).

[16] N. T. Le and A. Ritter. Are Large Language Models Robust Zero-shot Coreference Resolvers?, (2023).

[17] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end Neural Coreference Resolution. In 'Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing'. Association for Computational Linguistics, (2017).

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, (2019).

[19] J. Luoma and S. Pyysalo. Exploring cross-sentence contexts for named entity recognition with bert, (2020).

[20] L. v. d. Maaten and G. Hinton. *Visualizing data using t-sne.* Journal of Machine Learning Research **9** (2008), 2579–2605.

[21] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, (2018).

[22] W. Medhat, A. Hassan, and H. Korashy. *Sentiment analysis algorithms and applications: A survey.* Ain Shams Engineering Journal **5** (2014), 1093–1113.

[23] N. S. Moosavi and M. Strube. Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In 'Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', 632–642, Berlin, Germany, (August 2016). Association for Computational Linguistics.

[24] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, L, and R. Lowe. Training language models to follow instructions with human feedback, (2022).

[25] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In 'Joint Conference on EMNLP and CoNLL - Shared Task', 1–40, Jeju Island, Korea, (July 2012). Association for Computational Linguistics.

[26] A. Radford and K. Narasimhan. Improving language understanding by generative pre-training. (2018).

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. (2019).

[28] S. T. Roweis and L. K. Saul. *Nonlinear dimensionality reduction by locally linear embedding.* Science **290** (2000), 2323–2326.

[29] J. Straková, M. Straka, and J. Hajič. Neural architectures for nested ner through linearization, (2019).

[30] J. B. Tenenbaum, V. De Silva, and J. C. Langford. *A global geometric framework for nonlinear dimensionality reduction.* Science **290** (2000), 2319–2323.

[31] Q. Wang, Z. Mao, B. Wang, and L. Guo. *Knowledge Graph Embedding: A Survey of Approaches and Applications.* IEEE Transactions on Knowledge and Data Engineering **29** (2017), 2724–2743.

[32] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik. *Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization.* Journal of Machine Learning Research **22** (2021), 1–73.

[33] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, and et al. Ontonotes release 5.0, (2013).

[34] P. Xia, J. Sedoc, and B. V. Durme. Incremental Neural Coreference Resolution in Constant Memory, (2020).

[35] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. Luke: Deep contextualized entity representations with entity-aware self-attention, (2020).

[36] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, (2020).

[37] J. Zhanming and L. Wei. Dependency-Guided LSTM-CRF for Named Entity Recognition, (2019).

Understanding Neural Blind Image Deconvolution*

Antonie Brožová
brozoant@fjfi.cvut.cz

study programme: Mathematical Engineering
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. Blind image deconvolution is a severely ill-posed problem requiring a suitable regularization. Zero-shot deep learning algorithm SelfDeblur which is based on deep image prior efficiently recovers sharp images without any explicit regularization. This paper attempts to shed some light on possible reasons of its success, namely the deep image prior and the effect of the optimization method. An approach to the analysis of the results is also discussed with a focus on stochastic influences and differentiation between the no-blur solution and a solution with artifacts.

Keywords: blind image deconvolution, deep learning, deep image prior

Abstrakt. Slepá dekonvoluce obrazu je špatně podmíněná úloha vyžadující vhodnout regularizaci. Algoritmus SelfDeblur založený na hluboké neuronové síti sloužící jako deep image prior je schopný úspěšně zrekonstruovat rozmazaný obraz bez další explicitní regularizace a učení se na datasetu. Tento příspěvek se snaží vysvětlit možné důvody jeho efektivity, zaměřuje se především na roli deep image prior a metodu optimalizace. Dále je diskutována analýza výsledků vzhledem ke stochastickým vlivům a možnost odlišení různě poškozených řešení.

Klíčová slova: slepá dekonvoluce obrazu, hluboké učení, deep image prior

1 Introduction

Recovery of a sharp, clean image from a degraded one is a difficult task regardless of the type of degradation. This paper deals with blurring, which may be caused by a relative motion of a camera and a scene, turbulence in the atmosphere or the focus of a camera. 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 matrix. The deconvolution is basically an inverse operation to the convolution with the aim of recovering the sharp image from the blurred

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

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.

Minimizing (2) is difficult since there can be a high amount of local minima and apart from the ground-truth solution it is minimized globally by the trivial solution. This solution is called the no-blur solution and is not ruled out by the assumption that the sum of the elements of the PSF equals to one. Therefore, it is necessary to add some regularizer to (2) that helps recover the real sharp image $\boldsymbol{x}$.

1.1 BID methods

As the problem is highly ill-posed, some prior information is vital to the estimation. The Bayesian approach received a lot of attention at the beginning of the century, starting with Miskin and MacKay [10], Likas and Galatsanos [9] and Molina et al. [11]. Variational Bayes [17], [7] and Maximum Aposteriori (MAP) [8], [12] 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 estimates of the sharp images degraded by artifacts.

Another interesting approach to the problem of blind image deconvolution is deep learning [3], [21], [22]. Deep learning models usually require training on large datasets, giving them more information than the traditional methods get and, therefore, outperforming them. However, 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) [18] and they state that the structure of a deep neural network is a regularizer of the problem itself and that it may prefer images with certain characteristics. They successfully used it for image denoising, inpainting, and superresolution. Ren et al. combined the DIP image network with a fully connected network representing the PSF in 2020 and proposed SelfDeblur [13]. This model deblurs images without any training dataset and outperforms the traditional methods that are used for BID. A similar approach was chosen by Asim et al. in [1] exploiting the structure of generative networks in combination with classical handcrafted priors. DualDeblur [16] utilizes DIP and multiple blurry images. Wang et al. focus on the PSF and represent it with DIP as well as the sharp image [19]. Huo et al. proposed to combine DIP with variational Bayes in [4].

Reconstructions obtained with SelfDeblur are far from perfect. There are several issues that are connected to optimization as well as to the general ambiguity of BID. The deep learning nature of the algorithm does not simplify their analysis nor search for their solution. These topics will be discussed in this paper.

1.2 SelfDeblur

As described in [13], the model combines two generative neural networks, one for the sharp image, denoted as $\mathcal{G}_x$, and one for the PSF, denoted as $\mathcal{G}_k$. The estimates of the

sharp image and the PSF are generated by inputting fixed arrays $\boldsymbol{z}_x$ and $\boldsymbol{z}_k$, that are randomly sampled from uniform distribution, into the networks. The deconvolution is then formulated as

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

where $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_k$ represent learnable parameters of the neural networks. The restrictions on the values of the image and the PSF can be easily incorporated using softmax and sigmoid output layers. $\mathcal{G}_x$ is, as in [18], 5-level U-net [14] 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 optimized jointly in 5000 epochs using Adam optimizer [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 $\boldsymbol{\theta}_k$ (a bias vector if it was understood as a neural network) and a softmax output layer (denoted as $\sigma(.)$) because we mainly focus on image network behavior. It is optimized with a learning rate 10^{-2}. Apart from that, optimization tricks used in SelfDeblur - namely random perturbations of $\boldsymbol{z}_x$ and learning rate scheduling - are not used.

2 Metrics

The quality of a reconstruction of an image is commonly assessed by peak signal-to-noise ratio (PSNR). For two images $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, PSNR($\boldsymbol{x}_1$, $\boldsymbol{x}_2$) measures their similarity based on the ratio of maximal value and their MSE (power of corrupting noise). Denoting the estimate of the sharp image as $\boldsymbol{x}$ and ground-truth sharp image as $\boldsymbol{x}_{GT}$, there are two main reasons for PSNR($\boldsymbol{x}$, $\boldsymbol{x}_{GT}$) to be small: the estimate contains artifacts or the algorithm achieved the no-blur solution. To distinguish between these two, we propose to use not only PSNR-GT($\boldsymbol{x}$) := PSNR($\boldsymbol{x}$, $\boldsymbol{x}_{GT}$) but also PSNR-NB($\boldsymbol{x}$) := PSNR($\boldsymbol{x}$, $\boldsymbol{x}_{NB}$) measuring closeness to the no-blur solution $\boldsymbol{x}_{NB}$. Improvement in quality of an image can be measured by improved peak signal-to-noise ratio (ISNR) defined as ISNR($\boldsymbol{x}$) = PSNR-GT($\boldsymbol{x}$) - PSNR-NB($\boldsymbol{x}$) and is useful especially when results on more than one blurred image are compared.

Two loss functions will be used for analysis: BID loss measuring mean-squared error between the blurred image and convolution of estimates

$$\mathcal{L}_{BID}\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_x\right) = \frac{1}{nm} ||\boldsymbol{d} - \sigma\left(\boldsymbol{\theta}_k\right) \circledast \mathcal{G}_x\left(\boldsymbol{\theta}_x|\boldsymbol{z}_x\right)||_2, \tag{4}$$

and simpler U-net loss measuring mean-squared error between U-net output and an image $\boldsymbol{x}$

$$\mathcal{L}_{Unet}\left(\boldsymbol{\theta}_x|\,\boldsymbol{x}\right) = \frac{1}{nm} ||\boldsymbol{x} - \mathcal{G}_x\left(\boldsymbol{\theta}_x|\boldsymbol{z}_x\right)||_2. \tag{5}$$

3 Stochasticity and initialization

Before studying different aspects of SelfDeblur, it is worth noting that there are a couple of sources of stochasticity that make the analysis of its results a bit tricky. As was already mentioned, the input of the image network $\mathcal{G}_x$ is randomly sampled. Moreover, the network itself is initialized randomly using Kaiming uniform initialization scheme [5].

Another source is an implementation of certain functions that are used to construct the U-net: bilinear upsampling and convolution. While it is possible to use deterministic convolution in Python 3.8 with pytorch 2.0.1, bilinear upsampling needs to be replaced by some other operation to achieve reproducible results. Nearest neighbor upsampling was used to study the effects of the implementation of the U-net in this paper.

Firstly, to test the influence of a GPU, 100 runs of SelfDeblur with nondeterministic U-net were performed on NVIDIA GeForce RTX 2080 and NVIDIA TITAN V; one blurred image from the Levin dataset [8] was used. Graph (a) in Figure 1 shows that the obtained results differ with the type of GPU. Secondly, SelfDeblur was run 100 times with three combinations of initial values of parameters $\boldsymbol{\theta}_x$ and the input array $\boldsymbol{z}_x$ on NVIDIA GeForce RTX 2080. Apart from these nondeterministic runs, deterministic one was carried out for comparison. Graph (b) in Figure 1 shows that there is a strong dependency on initial values. Furthermore, deterministic computations may lead to a solution very different from the most likely one. On the right side of the figure, the three deterministically obtained deblurring results are depicted.

This analysis shows how sensitive the algorithm is to small changes - differences in computation caused by nondeterministic implementations of functions may lead to completely different results as well as random initialization. It also shows that the efficiency of an algorithm like SelfDeblur should not be judged only from one value, but histograms should be compared instead.

4 The prior

4.1 Deep image prior

Authors of DIP [18] state that the prior information helping recover the clean image is formed by a structure of the image network. This claim is supported only by experiments and it is not clear what is the key aspect leading the algorithm to the right solution. An important observation supporting the hypothesis is that naturally-looking images get learned faster by the U-net. This is illustrated in Figure 2, where 10 runs of simple U-net optimization (minimization of $\mathcal{L}_{Unet}$ loss function) were executed for ground-truth sharp image and blurred image from the Levin dataset [8], and an image with artifacts which was obtained by deblurring by a simpler method than SelfDeblur. According to this experiment, U-net prefers smoother images which generally achieve higher values of PSNR-GT than images with artifacts. Shi et al. [15] explain this behavior by the ability of convolutional networks to learn the lower frequency information faster than the higher frequency one. In the case of deblurring, early stopping is necessary to achieve a naturally looking smooth estimate. Unfortunately, the blurred image is the no-blur solution in BID, so this prior should not be strong enough to achieve the correct reconstruction.

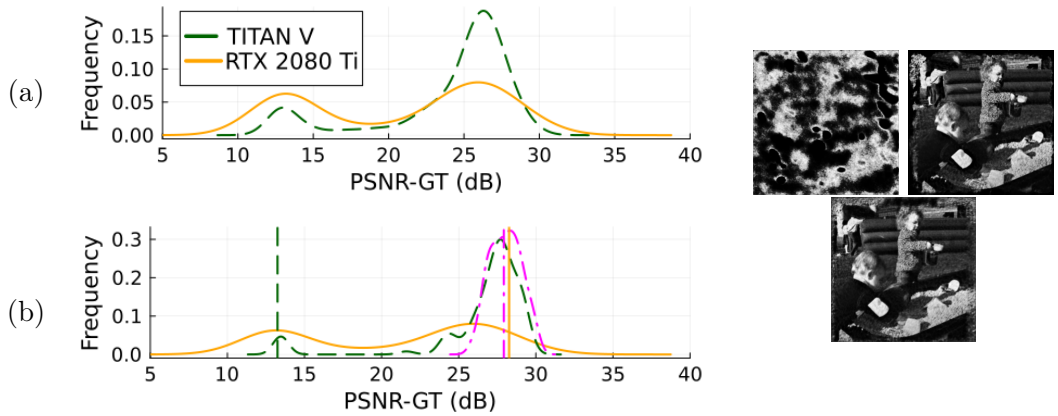


Figure 1: Comparison of image deblurring with different initial values and GPU setting. **(a)** PSNR-GT histograms of results obtained by nondeterministic computations on NVIDIA GeForce RTX 2080 Ti and NVIDIA TITAN V. **(b)** PSNR-GT histograms for three different initial values and nondeterministic computations on NVIDIA GeForce RTX 2080 Ti. The vertical lines show PSNR of a result obtained by deterministic operations from corresponding initial values. The three deterministically reconstructed images are displayed on the right side of the figure.

It also should be noted that the learning gets more unstable with lower loss value, which is one of the reasons of extremely low values of PSNR-GT in histograms. These cases should be excluded from the analysis if the value of loss is not reasonably low.

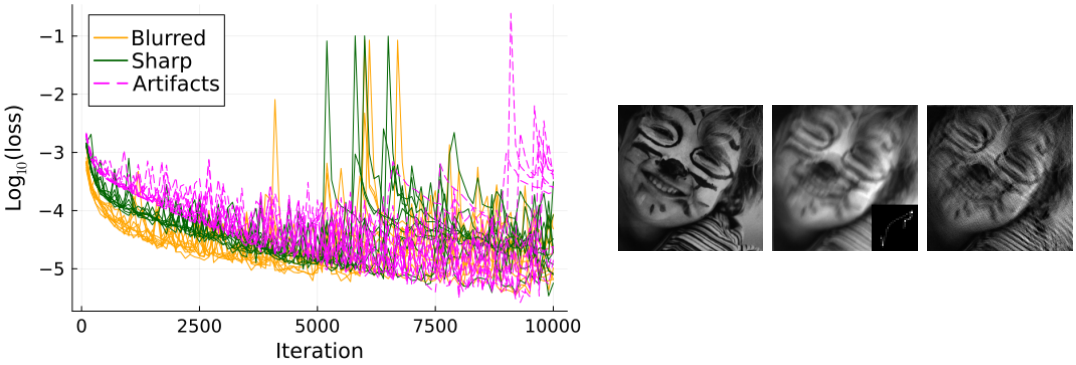


Figure 2: Comparison of the speed of learning of the sharp image, the blurred image, and an image with artifacts displayed on the right side of the figure.

4.2 Initialization of the PSF

Since DIP inside SelfDeblur should prefer the no-blur solution, learning of the PSF may be the reason why SelfDeblur achieves the correct sharp solution. To study how strongly

SelfDeblur is attracted to the no-blur and the ground-truth solutions, an augmented loss was created

$$\mathcal{L}\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_x | \, \boldsymbol{x}\right) = \alpha \mathcal{L}_{BID}\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_x\right) + (1-\alpha)\mathcal{L}_{Unet}\left(\boldsymbol{\theta}_x | \, \boldsymbol{x}\right),$$

for $\alpha \in \langle 0, 1 \rangle$. By setting the value of α and choosing $\boldsymbol{x}$ to be either $\boldsymbol{x}_{GT}$ or $\boldsymbol{x}_{NB}$ we can observe how quickly SelfDeblur learns the chosen image target and how PSF estimate changes. Figure 3 shows that for $\alpha = 0.1$ BID loss is minimized faster for the ground-truth target than for the no-blur one in the first 500 iterations. On the other hand, U-net loss is minimized faster for the no-blur target the whole time of the optimization. Both loss functions reach lower values for the no-blur target, which suggests that the descent of the BID loss is greater in the direction of the no-blur solution than in the direction of the ground truth solution.

PSF is in all our experiments initialized by a flat array and its evolution is depicted in Figure 3. For $\alpha = 0.9$, which gives most of the weight to BID loss, we can see that PSF estimates are similar after the first 100 iterations and they are closer to the ground-truth estimate in the case of the no-blur target. We, therefore, suggest that the initialization of the PSF helps to direct the optimization towards the ground-truth solution. Golatkar et al. [2] argue that the beginning of the optimization of a deep neural network is the key to achieving a good solution in terms of generalization. The effect of the PSF learning may be similar in this case, it helps the algorithm find the direction of a gradient towards the correct minima. As can be seen from Figure 3, the speed of learning slows down with the number of iterations and the estimated solution pair does not get significantly changed in the late steps of the algorithm, apart from the instability at the low values of the loss function.

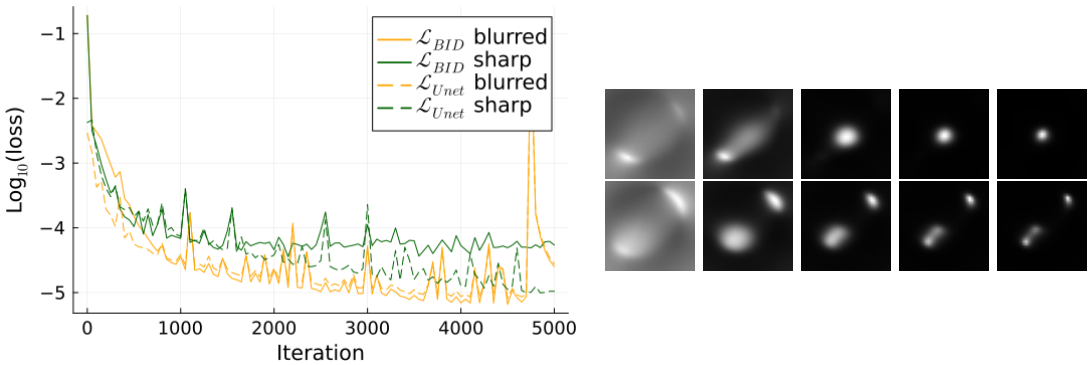


Figure 3: Speed of learning for $\mathcal{L}_{Unet}$ and $\mathcal{L}_{BID}$ and estimates of the PSF. Speed of learning is displayed for $\alpha = 0.1$. Estimates of PSF are displayed in iteration numbers 100, 200, 300, 400, and 500 for $\alpha = 0.9$. The first row shows the case of the no-blur target, the second row the case of the ground-truth target.

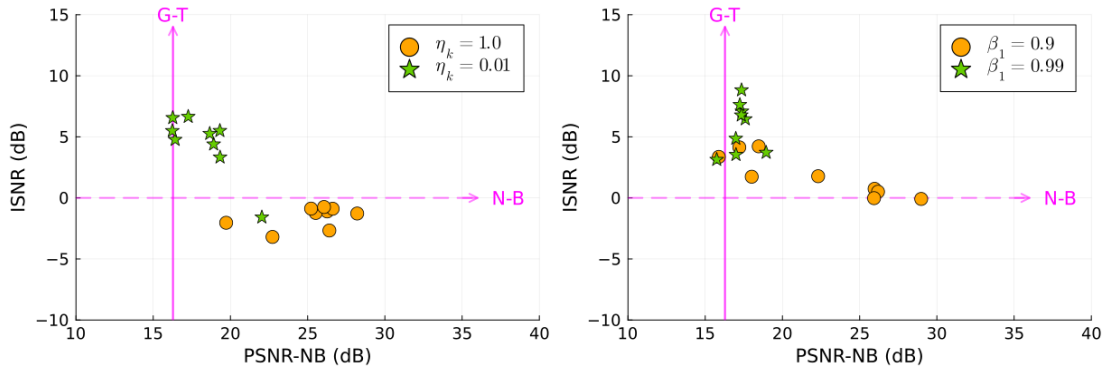


Figure 4: Sensitivity of the solution to the optimizer hyper-parameters in terms of PSNR-NB and ISNR. **Left**: variation of the PSF learning rate η_k. **Right**: variations of β_1 of the Adam optimiser of the U-net. Solid vertical line points in the direction of the ground-truth solution, dashed horizontal line in the direction of the no-blur solution.

4.3 Parameters of the optimization

The optimization setting plays an important role in directing the algorithm toward the correct solution. Two groups of parameters are learned $(\boldsymbol{\theta}_x, \boldsymbol{\theta}_k)$ and their interplay influences the solution. Both are optimized with the Adam optimizer [6], which has three hyperparameters: learning rate η, β_1, and β_2. Figure 4 and left graph in Figure 5 show how results of deblurring of one image from the Levin dataset [8] differ depending on the setting of the hyperparameters. Some combinations of learning rates of the sharp image and PSF prefer the correct solution while some the no-blur one. Apart from that, even the forgetting parameters inside of the Adam optimizer - in this case β_1 - direct the optimization toward different solutions. Unfortunately, every blurred image prefers a different setting of learning rates, this is illustrated on the right side of Figure 5, where there are results of deblurring of two different images. Moreover, the two depicted cases have the same ratio of η_k and η_x, so there seems to be no dependence on the ratio as well. This result makes it difficult to improve the optimization, the only option may be to use some meta-learning to estimate the right hyperparameters or even make it iteration-number dependent.

5 Conclusion

This paper focused on understanding neural blind image deconvolution, namely the effect of deep image prior and optimization method in SelfDeblur algorithm. It was shown that the deep image prior prefers naturally-looking smooth images, which in the case of the blind image deconvolution is not necessarily a good prior because the blurred image is smoother than the sharp one. The reason why SelfDeblur avoids the trivial no-blur solution may lie in the optimization method - hyperparameters of optimizer - and initialization of the PSF to a flat array. Unfortunately, the best choice of hyperparameters depends on the blurred image, so it is difficult to find a good setting. Apart from that, it was shown that the choice of GPU, nondeterministic implementation of convolution and

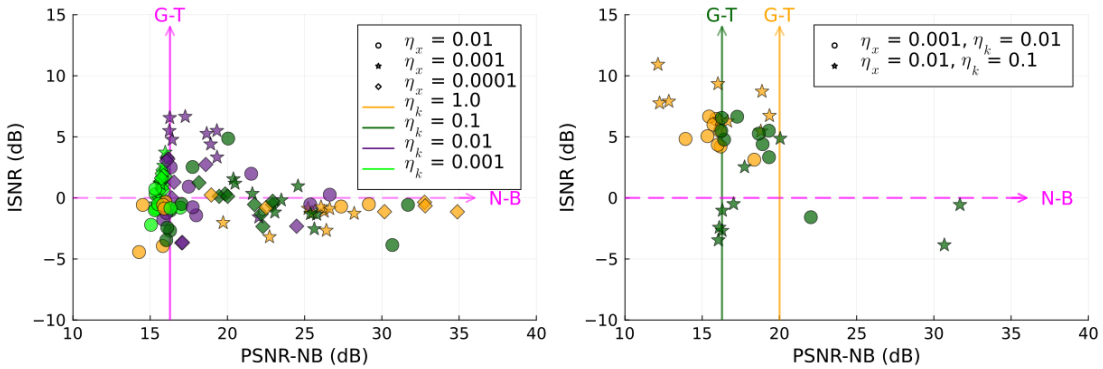


Figure 5: Sensitivity of the solution to the learning rates in terms of PSNR-NB and ISNR. **Left**: one image and different combinations of learning rates of PSF η_k and U-net η_x. **Right**: for two different images, each denoted by a separate color. Solid vertical line points in the direction of the ground-truth solution, dashed horizontal line in the direction of the no-blur solution.

bilinear upsampling, and initial values of the noise array and parameters of the U-net influence the deblurring results. New metric to differentiate between the ground-truth solution and the no-blur solution was proposed and figures in the paper illustrated its benefit in understanding the quality of a reconstruction.

References

[1] M. Asim, F. Shamshad, and A. Ahmed. *Blind image deconvolution using deep generative priors*. IEEE Transactions on Computational Imaging **6** (2020), 1493–1506.

[2] A. S. Golatkar, A. Achille, and S. Soatto. *Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence*. Advances in Neural Information Processing Systems **32** (2019).

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

[4] D. Huo, A. Masoumzadeh, R. Kushol, and Y.-H. Yang. Blind image deconvolution using variational deep image prior, (2023).

[5] H. Kaiming and et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In '2015 IEEE International Conference on Computer Vision (ICCV)', 1026–1034, Santiago, Chile, (2015). IEEE.

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

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

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

[16] C. J. Shin, T. B. Lee, and Y. S. Heo. *Dual image deblurring using deep image prior.* Electronics **10** (2021).

[17] 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.

[18] 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.

[19] Z. Wang, Z. Wang, Q. Li, and H. Bilen. Image deconvolution with deep image and kernel priors. In 'Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops', 0–0, (2019).

[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** (8 2022).

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

Which Graph Properties Affect GNN Performance for a Given Downstream Task?

Marek Dědič
dedicma2@fjfi.cvut.cz

study programme: Mathematical Engineering
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. Machine learning algorithms on graphs, in particular graph neural networks, became a popular framework for solving various tasks on graphs, attracting significant interest in the research community in recent years. As presented, however, these algorithms usually assume that the input graph is fixed and well-defined and do not consider the problem of constructing the graph for a given practical task. This work proposes a methodical way of linking graph properties with the performance of a GNN solving a given task on such graph via a surrogate regression model that is trained to predict the performance of the GNN from the properties of the graph dataset. Furthermore, the GNN model hyper-parameters are optionally added as additional features of the surrogate model and it is shown that this technique can be used to solve the practical problem of hyper-parameter tuning. We experimentally evaluate the importance of graph properties as features of the surrogate model with regards to the node classification task for several common graph datasets and discuss how these results can be used for graph composition tailored to the given task. Finally, our experiments indicate a significant gain in the proposed hyper-parameter tuning method compared to the reference grid-search method.

Keywords: Graph neural network, Graph properties, Meta-learning, Hyper-parameter optimization

Abstrakt. Algoritmy strojového učení na grafech, zejména grafové neuronové sítě, se staly populárním nástrojem pro řešení nejrůznějších úloh na grafech a v posledních letech přitahují značný zájem vědecké komunity. V prezentované podobě však tyto algoritmy obvykle předpokládají, že vstupní graf je pevně daný a dobře definovaný, a neuvažují problém konstrukce grafu pro danou aplikační úlohu. Tato práce navrhuje metodický způsob propojení vlastností grafu s efektivitou GNN řešící danou úlohu na daném grafu prostřednictvím náhradního regresního modelu, který je naučen předpovídat efektivitu GNN modelu z vlastností grafového datasetu. Hyperparametry modelu GNN mohou navíc být přidány jako další příznaky náhradního modelu a je ukázáno, že tuto techniku lze použít k řešení problému optimalizace hyperparametrů. Experimentálně vyhodnocujeme význam jednotlivých vlastností grafu jako příznaků náhradního modelu s ohledem na úlohu klasifikace uzlů pro několik běžných grafových datových sad a diskutujeme, jak lze tyto výsledky využít pro konstrukci grafu přizpůsobenou dané úloze. Naše experimenty ukazují na významný přínos navrhované metody ladění hyperparametrů ve srovnání s referenční grid-search metodou.

Klíčová slova: Grafová neuronová síť, Vlatnosti grafů, Meta-learning, Optimalizace hyperparametrů

Full paper: P. Procházka, M. Mareš, and M. Dědič. Which Graph Properties Affect GNN Performance for a Given Downstream Task? In 'Proceedings of the 23nd Conference Information Technologies – Applications and Theory (ITAT 2023)', volume 3498 of *CEUR Workshop Proceedings*, 58–66, Tatranské Matliare, Slovakia, (October 2023).

LQR-Trees with Sampling Based Exploration of the State Space*

Jiří Fejlek
fejlejir@fjfi.cvut.cz

study programme: Mathematical Engineering
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Stefan Ratschan, Department of Artificial Intelligence
Institute of Computer Science, CAS

Abstract. This paper introduces an extension of the LQR-tree algorithm [6, 5], which is a feedback-motion-planning algorithm for stabilizing a system of ordinary differential equations from a bounded set of initial conditions to a goal. The constructed policies are represented by a tree of exemplary system trajectories, so called demonstrations [4], stabilized via linear-quadratic regulator (LQR) feedback controllers. These stabilized demonstrations are used to cover the investigated set: Either a conservative region of attraction of the goal set can be computed for each demonstration [6], or these regions can be estimated via system simulations [5].

The key component of any LQR-tree algorithm is a demonstrator, a procedure that provides control inputs that steer the system into the goal set. In previous work [6, 5, 2], this demonstrator took the form of a trajectory optimization method [1]. However, such solvers require appropriate initial guesses to provide valid results. If these guesses are not good enough, the demonstrator regularly fails, causing the LQR-tree algorithm to progress slowly, if at all. This was the case in the implementation of LQR-trees in [5], where the demonstrator was initialized with failed system simulations generated during the run of the LQR-tree algorithm.

In this paper, we remedy this issue by extending the LQR-tree algorithm with exploration of the state space based on randomized motion-planning. More specifically, we use rapidly-exploring random trees (RRT) [3]. The RRT algorithm is an incremental algorithm that creates a tree in the state space by connecting new states to already explored ones. We use it to further extend the LQR-tree into areas of the state space that are not yet stabilized by the current LQR-tree. The newly discovered connections then serve as initial solutions to the original demonstrator based on a trajectory optimization method.

We provide computational experiments on several examples of dimension up to twelve that illustrate the practical applicability of the method and we compare our generation of demonstrations to the one in [5]. In this comparison, the exploring LQR-tree algorithm is more reliable and often significantly faster.

Keywords: nonlinear systems, motion planning, learning from demonstrations

Abstrakt. V tomto článku zavedeme rozšíření LQR-tree algoritmu [6, 5], což je algoritmus pro konstrukci zpětnovazebného řízení, které stabilizuje systém popsaný soustavou diferenciálních z omezené počáteční množiny do dané cílové množiny. Toto řízení je zadáno pomocí stromu exemplárních systémových trajektorií, tzv. demonstrací [4], stabilizovaných lineárně-kvadratickým

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

regulátorem (LQR). Tyto stabilizované demonstrace jsou použity k pokrytí počáteční množiny. Toto pokrytí může být určeno buď napočítáním konzervativních oblastí atrakce cílové množiny [6], anebo odhadnuto pomocí simulací [5].

Klíčovým prvkem jakéhokoliv LQR-tree algoritmu je demonstrátor, který napočítává řízení, které řídí systém do cílové množiny. V minulých pracích, byl tento demonstrátor implementován v podobě řešiče optimalizace trajektorií [1]. Takový řešič ale vyžaduje kvalitní počáteční odhady řešení aby poskytl validní demonstraci. Pokud tyto počáteční odhady nejsou dostatečně dobré, demonstrátor často selhává, což způsobuje pomalý, pokud vůbec nějaký, postup LQR-tree algoritmu. Takové chování LQR-tree algoritmu je pozorováno v [5], kde byl demonstrátor inicializován neúspěšnými simulacemi systému.

Abychom vyřešili tento problém, navrhujeme rozšíření LQR-tree algoritmu o exploraci stavového prostoru založené na znáhodněném plánování. Jmenovitě aplikujeme algoritmus RRT (rapidly-exploring random trees) [3]. Algoritmus RRT je inkrementální algoritmus, který vytváří strom ve stavovém prostoru aplikováním náhodně zvolených akcí z již prozkoumaných stavů systému. Toho využijeme pro rozšiřování stromu demonstrací do částí stavového prostoru, které ještě nejsou stabilizovány. Tyto nová napojení následně slouží jako počáteční řešení pro původní demonstrátor.

V článku poskytneme výpočetní experimenty na úlohách až do dimenze dvanáct, které ukazují praktické užití naší metody. Též porovnáme náš algoritmus s generováním demonstrací v algoritmu [5]. Ukážeme, že náš algoritmus je výrazně spolehlivější a často i rychlejší.

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

Full paper: Jiří Fejlek and Stefan Ratschan. LQR-trees with Sampling Based Exploration of the State Space, arXiv:2303.00553 (`https://arxiv.org/abs/2303.00553`, accepted to IROS 2023), 2023.

References

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

[2] J. Fejlek and S. Ratschan. *Computation of feedback control laws based on switched tracking of demonstrations.* Submitted, https://arxiv.org/abs/2011.12639 (2022).

[3] S. M. LaValle. *Rapidly-exploring random trees: A new tool for path planning.* Algorithmic Foundations of Robotics V (1998).

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

[5] 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.

[6] 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.

Butterfly Diffusion over Sparse Point Sets*

František Gašpar
frantisek.gaspar@fjfi.cvut.cz

study programme: Mathematical Engineering
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 graph-based random walk model of fractal diffusion is limited in its use to the connected sets and does not allow for direct fractal dimension estimation based on observed data. We discuss a task of directly obtaining accurate fractal dimension estimates and propose butterfly diffusion as an alternative approach using an explicit relation between walk and fractal dimensions. The validity of the presented approach is evaluated and statistical properties of the dimension estimates are presented. Through experiments on self-similar sets, we demonstrate the effectiveness of this approach in producing unbiased dimension estimates, offering a promising tool for fractal analysis and Monte Carlo simulations. The estimate of fractal dimension can be also created from spectral dimension, but this approach is less general and less accurate.

Keywords: dimension estimation, fractal dimension, point set, resistance scaling, walk dimension

Abstrakt. Model fraktální difúze založený na náhodné procházce grafem je omezen na souvislé množiny a neumožňuje přímý odhad fraktální dimenze na základě pozorovaných dat. Práce se věnuje úkolu přímého získání přesných odhadů fraktální dimenze a navrhujeme *butterfly diffusion* jako alternativní přístup využívající explicitní vztah mezi procházkovou a fraktálními dimenzí. Vyhodnocuje se platnost tohoto přístupu a uvádějí se statistické vlastnosti odhadů dimenzí. Prostřednictvím experimentů na soběpodobných množinách je demonstrována účinnost tohoto přístupu při vytváření nestranných odhadů dimenzí, který nabízí slibný nástroj pro fraktální analýzu a Monte Carlo simulace. Odhad fraktální dimenze lze také vytvořit z spektrální dimenze, tento přístup je však méně obecný a méně přesný.

Klíčová slova: odhad dimenze, fraktální dimenze, bodová množina, škálování odporu, procházková dimenze

Full paper: F. Gašpar and J. Kukal. *Butterfly Diffusion over Sparse Point Sets.* Under review in Physica A: Statistical Mechanics and its Applications.

*This work has been supported by the grant OP VVV MEYS, Czechia CZ.02.1.01/0.0/0.0/16019/0000765 as well as grant SGS23/190/OHK4/3T/14 grant of Czech Technical University in Prague

Non-local Relativistic δ-Shell Interactions

Lukáš Heriban
heribluk@cvut.cz

study programme: Mathematical Engineering
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Matěj Tušek, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This paper serves as the basic introduction to the problem of the Dirac operator with the singular potential. We try to emphasize the importance of dealing properly with formal operators, since the very similarly looking formal operators will correspond to the different mathematical models. Well known results concerning δ-interactions models are presented. New type of interactions is introduced and the self-adjointness of the operator is proved using the boundary triples technique.

Keywords: δ-interactions, boundary triples, Dirac operator, local shell interactions, non-local shell interactions, self-adjointness

Abstrakt. Tento článek slouží zejména jako úvod do problematiky singulárních poruch Diracova operátoru. Ukážeme, že podobné singulární potenciály vedou k různým matematickým modelům, a tudíž je důležité zacházet s formálními operátory opatrně. V článku jsou zopakované důležité výsledky pojednávající o lokální Diracově δ-interakci. Dále se čtenář může dočíst o nové nelokální δ-interakci. Nakonec je diskutována důležitá otázka samosdruženosti tohoto nového modelu.

Klíčová slova: δ interakce, Diracův operátor, hraniční trojce, lokální delta interakce, nelokální delta interakce, samosdruženost

1 Introduction

The δ-interactions model of physically relevant operators are important exactly solvable mathematical models which can serve as good approximations of real physical situation. We will turn our interest toward relativistic quantum model of the Dirac operator. One dimensional Dirac operator with the point interaction have been already studied in mathematical details in many articles. We must mention the famous article of Šeba [1] where he focused on two special cases of the point interaction of the Dirac operator and in some sense started the hunt for the singular perturbation of the Dirac operator. Five years after that, all self-adjoint realizations of the relativistic point interaction model in the first dimension were discovered in [19] by Benvegnu and Dabrowski. They started with the one dimensional Dirac operator restricted to the functions vanishing in the point of interaction

$$\begin{aligned} S\psi(x) &= -i\frac{\mathrm{d}}{\mathrm{d}x}\otimes\sigma_1\psi(x) + m\otimes\sigma_3\psi(x),\ x\in\mathbb{R}\setminus\{0\}\\ \operatorname{Dom} S &= \{\psi\in W^{1,2}(\mathbb{R};\mathbb{C}^2)\mid \psi(0)=0\}, \end{aligned}$$

where $m \in \mathbb{R}$ is used for the mass term, $\sigma_j, j \in \{1,2,3\}$ for the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

and by I_N we will denote the $N \times N$ identity matrix. Since the operator S is a symmetric operator with deficiency indices equal to two, they were able to use Von Neumann theory to find all self-adjoint extensions. We found out in [3] that almost all of these self-adjoint extensions correspond to the formal operator

$$D^A = -i\frac{\mathrm{d}}{\mathrm{d}x} \otimes \sigma_1 + m \otimes \sigma_3 + A\delta(x),$$

where δ stands for the Dirac delta function and A is 2×2 hermitian matrix. One can easily recover the proper definition of the operator D^A by examining the formal action and taking into account condition $D^A\psi \in L^2(\mathbb{R}; \mathbb{C}^2)$. From that we will end up with the operator

$$\begin{aligned} &(D^{\mathbb{A}}\psi)(x) = (\mathcal{D}\psi)(x),\ x \in \mathbb{R} \setminus \{0\}, \\ &\mathrm{Dom}(D^{\mathbb{A}}) = \{\varphi \in W^{1,2}(\mathbb{R} \setminus \{0\}) \otimes \mathbb{C}^2 \mid (2i - \sigma_1\mathbb{A})\varphi(0+) = (2i + \sigma_1\mathbb{A})\varphi(0-)\}. \end{aligned}$$

Natural question would be if this construction of the point interaction can be reproduced also in the higher dimensions. The easy answer is no. If the Dirac operator in dimension two or three is restricted to the functions vanishing in the one single point, we will end up with the essentially self-adjoint operator and therefore there is nothing interesting to be found. To introduce δ-interactions for those cases we need to work a little bit harder. Instead of perturbing the operator by the Dirac function we need to use a singular potential that would perturbed the operator on the higher dimensional subset of the Euclidean space. The potential we will consider is going to be single layer distribution which is sometimes called, in the context of δ-interactions, the δ-shell distribution. We will show that caution must be taken when handling with the formal operator with the single layer distribution since seemingly similar operators can correspond to self-adjoint Dirac operators with completely different transmission conditions along the boundary of the subset.

2 Local vs non-local singular potential

Let Ω be an open bounded simply connected subset of $\mathbb{R}^n$, $n = 2, 3$, with the Lipschitz smooth boundary Σ and $N = 2^{\lceil \frac{n}{2} \rceil}$. We will denote δ_Σ for the single layer distribution defined on $\mathcal{D}(\mathbb{R}^n; \mathbb{C}^N)$ as

$$(\delta_\Sigma, \varphi) = \int_\Sigma \varphi.$$

In the one dimensional situation there is no difference between projecting on and multiplying by δ-function since we have

$$|\delta\rangle\langle\delta|\varphi = (\delta, \varphi)\delta = \psi(0)\delta = \varphi\delta = \delta\varphi.$$

However, this is not the case for the single-layer distribution because we see that for $\varphi, \psi \in \mathcal{D}(\mathbb{R}^n; \mathbb{C}^N)$ we have

$$(\delta_\Sigma \varphi, \psi) = \int_\Sigma (\varphi\psi)$$

and

$$(|\delta_\Sigma\rangle\langle\delta_\Sigma|\varphi, \psi) = \int_\Sigma \varphi \int_\Sigma \psi,$$

which is clearly different. Therefore, we will need to differ between the local singular perturbation by the multiplication operator and the non-local singular perturbation by the projection. The Dirac operator with the potential $A\delta_\Sigma$, where A is a hermitian matrix valued function, can be described by a transmission condition along Σ and has been studied frequently, see, e.g. [8–11]. On the contrary, a formal operator $A|\delta_\Sigma\rangle\langle\delta_\Sigma|$ is not even formally symmetric for the non-constant matrix A. To introduce weighted and symmetric non-local singular perturbation we must be a little bit more shrewd. Instead of multiplying whole projection by one matrix valued function, we will multiply both bra and ket by $F, G \in L^2(\Sigma; \mathbb{C}^{N,N})$ respectively. Eventually, we have non-local δ-shell potential of a form

$$|F\delta_\Sigma\rangle\langle G\delta_\Sigma|, \tag{1}$$

where we naturally extend the definition of bra-vector as follows

$$|F\delta_\Sigma\rangle\langle G\delta_\Sigma|\varphi := F(\delta_\Sigma, G^*\varphi)\delta_\Sigma.$$

The potential (1) is clearly symmetrical if and only if in $L^2(\Sigma; \mathbb{C}^{N,N})$ we have

$$F \int_\Sigma G^*\varphi = G \int_\Sigma F^*\varphi.$$

3 Boundary triples

An investigation of the self-adjoitness can be quite challenging task. Fortunately, we have useful tool in the form of the quasi and the generalized boundary triples. In this section, we will summarize the most important definitions and theorems mainly following [7]. We encourage reader to study the theory in more details in [13–16]. In this section, we will assume that S is densely defined closed symmetric operator in a Hilbert space $\mathcal{H}$ and T stands for a linear operator satisfying $\overline{T} = S^*$. For our latter purposes S will be a restriction of the free Dirac operator D_0 to functions vanishing along Σ and the extension of S will be identified with the Dirac operators with the δ-shell potential.

Definition 1. *Let T be such that $\overline{T} = S^*$. A triple $(\mathcal{G}, \Gamma_0, \Gamma_1)$ consisting of a Hilbert space $\mathcal{G}$ and linear mappings $\Gamma_0, \Gamma_1 : \operatorname{Dom} T \to \mathcal{G}$ is called a quasi boundary triple for S^* if the following holds:*

1. *For all $f, g \in \operatorname{Dom} T$, $\langle Tf, g\rangle_\mathcal{H} - \langle f, Tg\rangle_\mathcal{H} = \langle \Gamma_1 f, \Gamma_0 g\rangle_\mathcal{G} - \langle \Gamma_0 f, \Gamma_1 g\rangle_\mathcal{G}$.*
2. *The range of $\Gamma = (\Gamma_0, \Gamma_1)$ is dense in $\mathcal{G} \times \mathcal{G}$.*

3. *The restriction* $T_0 := T \restriction \operatorname{Ker}\Gamma_0$ *is a self-adjoint operator in* $\mathcal{H}$.

If conditions (1) and (3) hold, and the mapping $\Gamma_0 : \operatorname{Dom} T \to \mathcal{G}$ is surjective, then $(\mathcal{G}, \Gamma_0, \Gamma_1)$ is called generalized boundary triple. Note that every generalized boundary triple is also a quasi boundary triple, because (2) follows from the defining properties of the generalized triple [16, Lemma 6.1].

Definition 2. *Let* S, T *be as above,* $(\mathcal{G}, \Gamma_0, \Gamma_1)$ *be a quasi boundary triple for* S^*, *and* $T_0 = T \restriction \operatorname{Ker}\Gamma_0$. *Then the associated* γ*-field and the Weyl function* M *are defined by*

$$\rho(T_0) \ni z \mapsto \gamma(z) = (\Gamma_0 \restriction \operatorname{Ker}(T - z))^{-1}$$

and

$$\rho(T_0) \ni z \mapsto M(z) = \Gamma_1(\Gamma_0 \restriction \operatorname{Ker}(T - z))^{-1}.$$

For a linear operator B in $\mathcal{G}$, we put

$$T_B = T \restriction \operatorname{Ker}(\Gamma_0 + B\Gamma_1). \tag{2}$$

Since $\operatorname{Dom} S = \ker\Gamma_0 \cap \ker\Gamma_1$ by [14, Prop. 2.2], $S \subset T_B$. The following theorem yields an eigenvalue condition for T_B, an alternative description of $\operatorname{Ran}(T_B - z)$, which may be used in the proof of self-adjointness of T_B, and a Krein-like formula for the resolvent of T_B.

Theorem 3. *Let* S, T *be as above,* $(\mathcal{G}, \Gamma_0, \Gamma_1)$ *be a quasi boundary triple for* S^*, $T_0 = T \restriction \operatorname{Ker}\Gamma_0$, *and* γ *and* M *denote the associated* γ*-field and the Weyl function, respectively. Finally, let* T_B *be given by* (2). *Then the following holds for all* $z \in \rho(T_0)$*:*

1. $z \in \sigma_{\mathrm{p}}(T_B)$ *if and only if* $0 \in \sigma_{\mathrm{p}}(I + BM(z))$. *Moreover,*

$$\operatorname{Ker}(T_B - z) = \{\gamma(z)\psi \mid \psi \in \operatorname{Ker}(I + BM(z))\}.$$

2. *If* $z \notin \sigma_{\mathrm{p}}(T_B)$, *then* $g \in \operatorname{Ran}(T_B - z)$ *if and only if* $B\gamma(\overline{z})^* g \in \operatorname{Ran}(I + BM(z))$.

3. *If* $z \notin \sigma_{\mathrm{p}}(T_B)$, *then*

$$(T_B - z)^{-1} g = (T_0 - z)^{-1} g - \gamma(z)(I + BM(z))^{-1} B\gamma(\overline{z})^* g \tag{3}$$

holds for all $g \in \operatorname{Ran}(T_B - z)$.

4 Relativistic δ-shell interactions

The time has come to introduce δ-shell interactions in dimensions two and three. To make everything cleaner, we will slightly abuse bra-ket notation even further in the following manner. The symbol $\langle \cdot, \cdot \rangle$ will be used for the scalar product, linear in the second argument, on $L^2(\mathbb{R}^n; \mathbb{C}^N)$ and we will naturally extend the notation to

$$\langle F, \varphi \rangle = \int_{\mathbb{R}^n} F^*(x)\varphi(x)\,\mathrm{d}x,$$
$$\langle F, G \rangle = \int_{\mathbb{R}^n} F^*(x)G(x)\,\mathrm{d}x,$$

where $\varphi \in L^2(\mathbb{R}^n;\mathbb{C}^N)$ and F, $G \in L^2(\mathbb{R}^n;\mathbb{C}^{N,N})$. Following the same pattern, we define finite rank linear operator $|F\rangle\langle G|$ in $L^2(\mathbb{R}^n;\mathbb{C}^N)$ as

$$|F\rangle\langle G|\varphi := F\langle G,\varphi\rangle = F\int_{\mathbb{R}^n} G^*(x)\varphi(x)\,\mathrm{d}x.$$

In the case of the integral operator K we will use $K(x,y)$ for its integral kernel. For the square root of the complex number $z \in \mathbb{C}\setminus(-\infty,0]$ we adopt the convention $\mathrm{Im}\sqrt{z} \geq 0$.

Let us now recall the free Dirac operator. For the two dimensional situation, we put

$$\alpha_1 = \sigma_1, \alpha_2 = \sigma_2, \alpha_0 = \sigma_3$$

and for the three dimensional case

$$\forall k \in \{1,2,3\}, \alpha_k = \begin{pmatrix} 0 & \sigma_k \\ \sigma_k & 0 \end{pmatrix}, \alpha_0 = \begin{pmatrix} I_2 & 0 \\ 0 & -I_2 \end{pmatrix}.$$

The Dirac operator D_0 acts like the following differential operator

$$\mathcal{D}_0 := -i(\alpha\cdot\nabla) + m\alpha_0$$

and is defined as

$$\begin{aligned} &D_0\varphi = \mathcal{D}_0\varphi, \\ &\mathrm{Dom}(D_0) = H^1(\mathbb{R}^n;\mathbb{C}^N). \end{aligned}$$

It is well known that the operator D_0 is self-adjoint, its spectrum is absolutely continuous and consists of

$$\sigma(D_0) = \sigma_{ac}(D_0) = (-\infty,-|m|]\cup[|m|,+\infty),$$

these results can be found for example in [12]. The resolvent for $z \in \mathbb{C}\setminus\sigma(D_0)$ is given by the integral operator $(D_0 - z)^{-1}(x,y) = R_z(x-y)$ where

$$R_z(x) = \frac{\sqrt{z^2-m^2}}{2\pi}K_1(-i\sqrt{z^2-m^2}|x|)\frac{\alpha\cdot x}{|x|} + \frac{1}{2\pi}K_0(-i\sqrt{z^2-m^2}|x|)(zI_2 + m\alpha_0)$$

for the second dimension, with K_j denoted for the modified Bessel functions of the second kind, and for the third dimension

$$R_z(x) = \left(zI_4 + m\alpha_0 + (1 - i\sqrt{z^2-m^2}|x|)\frac{i(\alpha\cdot x)}{|x|^2}\right)\frac{1}{4\pi|x|}\mathrm{e}^{i\sqrt{z^2-m^2}|x|}.$$

4.1 Local relativistic δ-shell interactions

Recall that we assume Σ to be the Lipschitz smooth boundary of an open bounded simply connected set $\Omega \equiv \Omega_+ \subset \mathbb{R}^n$, $n = 2, 3$. Denote *the outer domain* $\mathbb{R}^n \setminus \overline{\Omega_+}$ by Ω_-. Then we have decomposition of the Euclidean space as the disjoint union $\mathbb{R}^n = \Omega_+ \cup \Sigma \cup \Omega_-$.

Furthermore, we denote by $n(x)$ the unit normal vector at $x \in \Sigma$ pointing outwards of Ω_+. For $s \in [0,1]$, define the space

$$H^s_\alpha(\Omega_\pm) := \{\psi_\pm \in H^s(\Omega_\pm;\mathbb{C}^N) \mid (\alpha\cdot\nabla)\psi_\pm \in L^2(\Omega_\pm;\mathbb{C}^N)\}.$$

It was shown in [7, Lemma 4.1 and Corollary 4.6] that $\psi \in H^s_\alpha(\Omega_\pm)$ admit Dirichlet traces $\mathcal{T}_\pm$ in $H^{s-\frac{1}{2}}(\Sigma;\mathbb{C}^N)$. In particular, $\mathcal{T}_\pm\psi_\pm \in L^2(\Sigma;\mathbb{C}^N)$ for $\psi_\pm \in H^{\frac{1}{2}}_\alpha(\Omega_\pm)$.

Then, we have closed, symmetric and densely defined operator $S = D_0 \restriction H^1_0(\mathbb{R}^n \setminus \Sigma;\mathbb{C}^N)$ for which S^* is given by

$$\begin{aligned} &S^*(\psi_- \oplus \psi_+) = \mathcal{D}_0\psi_- \oplus \mathcal{D}_0\psi_+,\\ &\operatorname{Dom}(S^*) = \{\psi_- \oplus \psi_+ \mid \psi_\pm \in H^0_\alpha(\Omega_\pm)\},\end{aligned}$$

cf. [17, Prop. 3.1]. Choosing

$$T := S^* \restriction H^{\frac{1}{2}}_\alpha(\Omega_-) \oplus H^{\frac{1}{2}}_\alpha(\Omega_+), \tag{4}$$

the triple $(\mathcal{G},\Gamma_0,\Gamma_1)$, where $\mathcal{G} = L^2(\Sigma;\mathbb{C}^N)$ and

$$\begin{aligned} \Gamma_0\psi &= i(\alpha\cdot n)(\mathcal{T}_+\psi_+ - \mathcal{T}_-\psi_-) : \operatorname{Dom} T \to L^2(\Sigma;\mathbb{C}^N),\\ \Gamma_1\psi &= \frac{1}{2}(\mathcal{T}_+\psi_+ + \mathcal{T}_-\psi_-) : \operatorname{Dom} T \to L^2(\Sigma;\mathbb{C}^N),\end{aligned} \tag{5}$$

form a generalized boundary triple for S^*. Corresponding γ-field and Weyl function were shown to be

$$\forall x \in \mathbb{R}^n \setminus \Sigma, \quad \gamma(z)\psi(x) = \int_\Sigma R_z(x-y)\psi(y)\,\mathrm{d}\sigma(y)$$

and

$$\forall x \in \Sigma, \quad M(z)\psi(x) = \lim_{\rho\to 0}\int_{\Sigma\setminus B(x,\rho)} R_z(x-y)\psi(y)\,\mathrm{d}\sigma(y),$$

respectively. Here, $\gamma(z)$ is bounded and everywhere defined operator from $L^2(\Sigma;\mathbb{C}^N)$ to $L^2(\mathbb{R}^n;\mathbb{C}^N)$ with a compact adjoint and $M(z)$ is bounded and everywhere defined operator in $L^2(\Sigma;\mathbb{C}^N)$.

To recover proper definition of the Dirac operator perturbed by the formal potential $A\delta_\Sigma$. We will proceed similarly to the one dimensional situation. Firstly, one can verify, using integration by parts, following identity for $\psi = \psi_- \oplus \psi_+ \in \operatorname{Dom} T$

$$\mathcal{D}_0(\psi_- \oplus \psi_+) = T(\psi_- \oplus \psi_+) + i(\alpha\cdot n)(\mathcal{T}_+\psi_+ - \mathcal{T}_-\psi_-)\delta_\Sigma. \tag{6}$$

Next, we naturally extend the definition of $A\delta_\Sigma$ for $\psi \in \operatorname{Dom} T$ in the following way

$$(A\delta_\Sigma\psi, f) = \frac{1}{2}\int_\Sigma (\mathcal{T}_+\psi_+ + \mathcal{T}_-\psi_-)f, \quad f \in \mathcal{D}(\mathbb{R}^n;\mathbb{C}^N).$$

The condition $\psi \in \operatorname{Dom} T, (\mathcal{D}_0 + A\delta_\Sigma)\psi \in L^2(\mathbb{R}^n;\mathbb{C}^N)$ is true if and only if

$$i(\alpha\cdot n)(\mathcal{T}_+\psi_+ - \mathcal{T}_-\psi_-) + \frac{1}{2}A(\mathcal{T}_+\psi_+ - \mathcal{T}_-\psi_-) = 0. \tag{7}$$

These considerations allow us to define the following operator

Definition 4. *We define the linear operator D_A in $L^2(\mathbb{R}^n;\mathbb{C}^N)$ as follows*

$$D_A(\psi_- \oplus \psi_+) = \mathcal{D}_0\psi_- \oplus \mathcal{D}_0\psi_+$$
$$\operatorname{Dom} D_A = \{\psi_- \oplus \psi_+ \in H^{\frac{1}{2}}_\alpha(\Omega_-) \oplus H^{\frac{1}{2}}_\alpha(\Omega_+) \mid \psi \text{ satisfies (7)}\}.$$

We will call the operator D_A the Dirac operator with local δ-shell interaction.

It was proved that for the large family of hermitian-matrix valued functions A the operator D_A is indeed the self-adjoint extension of the operator S. For more details see [7]. We will try to show general ideas of the proof of the self-adjointness in the following section where we will introduce completely new type of interactions.

4.2 Non-local relativistic δ-shell interactions

We will go back and try to use (1) as our formal perturbation of the free Dirac operator. Embedding (6) and an extension of (1) to discontinuous functions along Σ in the following way

$$|F\delta_\Sigma\rangle\langle G\delta_\Sigma|\psi := F\int_\Sigma \left(G^*\frac{1}{2}(\mathcal{T}_+\psi_+ + \mathcal{T}_-\psi_-)\right)\delta_\Sigma.$$

we again use the condition $\forall\psi \in \operatorname{Dom} T, (\mathcal{D}_0 + |F\delta_\Sigma\rangle\langle\delta_\Sigma|)\psi \in L^2(\mathbb{R}^n;\mathbb{C}^N)$ to recover the transmission condition

$$i(\alpha\cdot n)(\mathcal{T}_+\psi_+ - \mathcal{T}_-\psi_-) + \frac{1}{2}F\int_\Sigma G^*(\mathcal{T}_+\psi + +\mathcal{T}_-\psi_-) = 0. \tag{8}$$

From that we define the following operator.

Definition 5. *By the Dirac operator with non-local δ-shell interaction of the type $|F\delta_\Sigma\rangle\langle G\delta_\Sigma|$ we mean the linear operator $D_{F,G}$ in $L^2(\mathbb{R}^n;\mathbb{C}^N)$ given by*

$$\operatorname{Dom} D_{F,G} = \{\psi_- \oplus \psi_+ \in H^{\frac{1}{2}}_\alpha(\Omega_-) \oplus H^{\frac{1}{2}}_\alpha(\Omega_+) \mid \psi \text{ satisfies (8)}\},$$
$$D_{F,G}(\psi_- \oplus \psi_+) = \mathcal{D}_0\psi_- \oplus \mathcal{D}_0\psi_+.$$

The transmission condition (8) may be rewritten as

$$\Gamma_0\psi + B\Gamma_1\psi = 0 \quad \text{with} \quad B = |F\rangle_\Sigma\langle G|_\Sigma, \tag{9}$$

where $|F\rangle_\Sigma\langle G|_\Sigma$ defined by

$$\varphi \mapsto F\int_\Sigma (G^*\varphi)$$

is a finite rank operator in $L^2(\Sigma;\mathbb{C}^N)$. Note that

$$(|F\rangle_\Sigma\langle G|_\Sigma)^* = |G\rangle_\Sigma\langle F|_\Sigma. \tag{10}$$

With this choice of B, $D_{F,G} = T_B$. In particular, $D_{0,0} = T_0 = D_0$ is the free Dirac operator. Hence, we may use Theorem 3 to show self-adjoitness of $D_{F,G}$ efficiently.

Theorem 6. *Let F, $G \in L^2(\Sigma; \mathbb{C}^{N,N})$ be such that*

$$|F\rangle_\Sigma\langle G|_\Sigma = |G\rangle_\Sigma\langle F|_\Sigma. \tag{11}$$

Then $D_{F,G}$ is a self-adjoint operator.

Proof. In view of (10), the condition (11) is equivalent to the hermiticity of $B = |F\rangle_\Sigma\langle G|_\Sigma$. The property (1) from Definition 1 together with (9) then imply that the operator $D_{F,G}$ is symmetric. Hence, to prove the self-adjointness of $D_{F,G}$ it is sufficent to show that $\forall z \in \mathbb{C} \setminus \mathbb{R}$, $\mathrm{Ran}(D_{F,G} - z) = L^2(\mathbb{R}^n; \mathbb{C}^N)$. By the symmetry of $D_{F,G}$, we also have $\sigma_\mathrm{p}(D_{F,G}) \subset \mathbb{R}$. Therefore, from the point (1) of Theorem 3, the operator $(I + BM(z))$ is injective for all $z \in \mathbb{C} \setminus \mathbb{R}$. Furthermore, B is a finite rank operator and thus compact. In addition, $M(z)$ is bounded, and so we deduce that the operator $BM(z)$ is also compact in $L^2(\Sigma; \mathbb{C}^N)$. On top of that, I is Fredholm operator with index 0 and the same holds true for its compact perturbation $(I + BM(z))$ which implies that the operator $(I + BM(z))$ is also surjective. This yields $\mathrm{Ran}(D_{F,G} - z) = L^2(\mathbb{R}^n; \mathbb{C}^N)$, due to the point (2) of Theorem 3. □

5 Conclusion

We looked at the already known results concerning the Dirac operator with the singular potential such as δ-function for the one dimension and the single layer distribution for higher dimensions. Also, completely new type of interaction, namely non-local δ-shell interactions, were introduced and the condition on the self-adjointness was found.

Another important result that can be discussed is the question of the existence of the regular approximation. Even though, the self-adjointness is important for the quantum theory, without the regular approximation one cannot hope to connect these mathematical models with the real physical situation. Another problem one can try to attack is the non-relativistic limit, to find the non-relativistic counterpart of the model. This seems to be particularly non-intuitive for the local δ-shell interaction, see [20], and even worse for the non-local situation.

References

[1] P. Šeba, *Klein's Paradox and the Relativistic Point Interaction.* Letters in Mathematical Physics 18, 1989, 77-86.

[2] M. Tušek, *Approximation of one-dimensional relativistic point interactions by regular potentials revised.* Lett. Math. Phys. 110, 2020.

[3] L. Heriban, M. Tušek, *Non-self-adjoint relativistic point interaction in one dimension.* J. Math. Anal. Appl. 516, 2022.

[4] B. Cassano, V. Lotoreichik, A. Mas, M. Tušek, *General δ-shell interactions for two-dimensional Dirac operator: self-adjointness and approximation.* Matématica Iberoamericana, 2022.

[5] J. Behrndt, M. Holzmann, M. Tušek, *Two-dimensional Dirac operator with general δ-shell interaction supported on a straight line.* J. Phys. A56, 2023.

[6] A. Mas, F. Pizzichillo, *Klein's paradox and the relativistic δ-shell interaction in* $\mathbb{R}^3$*.* Anal. and PDE 11, 2018.

[7] J. Behrndt, M. Holzmann, C. Stelzer, G. Stenzel, *Boundary triples and Weyl functions for Dirac operators with singular interaction.* arXiv:2211.05191, 2022.

[8] J. Behrndt, M. Holzmann, T. Ourmieres-Bonafas, K. Pankrashkin, *Two-dimensional Dirac operators with singular interactions supported on closed curves.* Journal of Functional Analysis 279, 2020.

[9] N. Arrizabalaga, A. Mas, L. Vega, *Shell interactions for Dirac operators.* J. Math. Pures Appl. (9) 102(4), 2014.

[10] J. Behrndt, P. Exner, M. Holzmann, V. Lotoreichik, *On Dirac operators in* $\mathbb{R}^3$ *with electrostatic and Lorentz scalar δ-shell interactions.* Quantum Studies 6, 2019.

[11] B. Benhellal, *Spectral properties of the Dirac operator coupled with δ-shell interactions.* Letters in Mathematical Physics 112(6), 2022.

[12] B. Thaller, *The Dirac equation*, Texts and Monographs in Physics, Springer-Verlag, Berlin, 1992.

[13] J. Behrndt, M. Langer, *Elliptic operators, Dirichlet-to-Neumann maps and quasi boundary triples*, in Operator methods for boundary value problems, volume 404 of London Math.Soc. Lecture Note Ser., 121–160. Cambridge Univ. Press, Cambridge, 2012.

[14] J. Behrndt, M. Langer, *Boundary value problems for elliptic partial differential operators on bounded domains*, J. Funct. Anal. 243(2), 2007.

[15] V. Derkach, M. Malamud, *Generalized resolvents and the boundary value problems for Hermitian operators with gaps*, J. Funct. Anal. 95(1), 1991.

[16] V. Derkach, M. Malamud, *The extension theory of Hermitian operators and the moment problem*, J. Math. Sci. 73(2), 1995.

[17] J. Behrndt, M. Holzmann, *On Dirac operators with electrostatic δ-shell interactions of critical strength*, J. Spectral Theory 10, 2020.

[18] M. Sh. Birman, M.Z. Solomjak, *Spectral Theory of Self-Adjoint Operators in Hilbert Spaces*, D. Reidel Publishing Co., Dordrecht, 1987.

[19] S. Benvegnu, L. Dabrowski, *Relativistic point interaction*, Letters in Mathematical Physics 30, 1994.

[20] J. Behrndt, M. Holzmann, G. Stenzel, *Schrödinger Operators with Oblique Transmission Conditions in* $\mathbb{R}^2$, Commun. Math. Phys. 401, 2023.

Self-Attention for Image Completion Task on the Calorimeter Data in High Energy Physics*

Kristina Jarůšková
jaruskri@fjfi.cvut.cz

study programme: Applied Informatics
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Quang Van Tran, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Research in the field of high energy physics relies significantly on artificial simulations of particle behaviour in the detector. Standard simulation tools such as Geant4 are based on Monte Carlo algorithms and provide high-fidelity simulations. However, they require vast computation resources, especially for the highly granular calorimeter part of the detector.

Generative deep learning algorithms offer a speed up of a few orders of magnitude for the price of a small loss of accuracy. Most of these models are based on MLPs or CNNs that are used to build an adversarial generative network or a variational autoencoder. However, each model is trained on a very specific detector and particle setting. In case of a small alteration to this setting, new model needs to be trained.

Transformer is an encoder-decoder model that is based on the attention mechanism. It is the state-of-the art approach in natural language processing due to its large learning capacity and the ability to adapt to different tasks. Hence, this approach seems to be a suitable candidate for a more general calorimeter simulation model. Unlike the convolution layers, the multi-head attention blocks do not introduce strong inductive bias to the model which enables learning complex dependencies within the data. Moreover, the training of the model is relatively fast thanks to the parallel nature of the multi-head attention computation. Both of these properties can be leveraged in the high energy physics domain.

In this work, we tested a transformer-inspired model on an auxiliary task of image completion. The calorimeter data are handled as images, split into smaller segments, partially masked, and processed by the model containing the self-attention blocks that restores the masked segments. We examine the influence of different pre-processing to the resulting reconstruction of images. We also experiment with the use of a graph neural network to obtain a better representation of the image segments.

Keywords: attention mechanism, calorimeter simulation, high energy physics, masked language modeling, transformer

Abstrakt. Simulace chování částic v detektoru jsou nezbytnou součástí výzkumu v oblasti částicové fyziky. Pro získání velmi přesných simulací jsou standardně používány nástroje jako Geant4, které jsou založené na Monte Carlo algoritmech. Tyto nástroje jsou však velmi výpočetně náročné, a to především v případě simulování odezvy kalorimetru, jednoho z obvyklých segmentů detektoru.

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS23/190/OHK4/3T/14.

Generativní deep learning modely dokáží simulovat odezvu kalorimetru na částice až o několik řádů rychleji za cenu jisté ztráty původní přesnosti. Většina z dnes užívaných modelů je založená na vícevrstvých perceptronech nebo konvolučních sítích, které jsou spojovány do komplexnějších architektur typu generativní kompetitivní síť nebo variační autoencoder. Zpravidla je však takový model natrénován pro velmi specifický typ detektoru a částice. Pokud jsou parametry detektoru nebo vstupní částice změněny, je potřeba model natrénovat znovu na odpovídajících datech.

Transformer je model typu encoder-decoder, jehož základním stavebním blokem je tzv. attention mechanismus. Jedná se o state-of-the-art přístup v oblasti zpracování textu. Jeho hlavní předností je velká kapacita z hlediska extrakce informací z dat a možnost adaptace na různé úlohy. Proto je tento typ modelu vhodný kandidátem pro sestavení obecnějšího nástroje pro simulaci kalorimetrů. Mechanismus multi-head attention nevnáší do modelu implicitní předpoklady o datech a lze jej díky tomu natrénovat i na datech s komplexní vnitřní strukturou. Výpočet multi-head attention lze navíc velmi dobře paralelizovat, což umožňuje relativně rychlé trénování modelu. Obě tyto vlastnosti jsou pro použití v částicové fyzice žádoucí.

V této práci je prezentován model inspirovaný transformerem, který byl použit pro pomocnou úlohu doplnění obrazu. Na trénovací data odezvy kalorimetru na částici se díváme jako na 3D obrázky, které na začátku rozdělíme na menší segmenty. Na většinu těchto segmentů následně aplikujeme masku a upravený obrázek zpracujeme modelem se self-attention bloky, který překryté segmenty zrekonstruuje a obrázek tak doplní. Navíc byl zkoumán vliv předzpracování dat na trénování modelu a úspěšnost rekonstrukce obrázků. Bylo také testováno použití grafové konvoluční sítě pro zlepšení embeddingu segmentů.

Klíčová slova: attention mechanismus, simulace kalorimetru, fyzika vysokých energií, maskované jazykové modelování, transformer

Full paper: K. Jaruskova and S. Vallecorsa. *Self-attention for Image Completion Task on the Calorimeter Data in High Energy Physics.* Full version to be published in the Proceedings of the 14th International Conference on Stochastic and Physical Monitoring Systems 2023.

An Improved Branch and Bound Algorithm for Phase Stability Testing of Multicomponent Mixture*

Martin Jex
jexmarti@fjfi.cvut.cz

study programme: Mathematical Engineering
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 Mikyška, J., Firoozabadi A., 2012, Investigation of Mixture Stability at Given Volume, temperature and number of moles, Fluid Phase Equilibria and the branch and bound algorithm from Smejkal T., Mikyška J., 2020, VTN-phase stability testing using the Branch and Bound strategy and the convex-concave splitting of Helmhotz free energy density, Fluid Phase Equilibria. In this contribution, we improve the algorithm 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 pressure function given by 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 Mikyška, J., Firoozabadi A., 2012, Investigation of Mixture Stability at Given Volume, temperature and number of moles, Fluid Phase Equilibria a metodě větví a mezí z Smejkal T., Mikyška J., 2020, VTN-phase stability testing using the Branch and Bound strategy and the convex-concave splitting of Helmhotz free energy density, Fluid Phase Equilibria. V tomto příspěvku zlepšujeme algoritmus o lepší zamítaní neperspektivních 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

*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.

rovnicí) přípustný. Pokud tomu tak není, jsme oprávněni tuto část přípustné množiny vyřadit z hledání. Tlak daný Pengovou-Robinsonovou stavovou rovnicí není konvexní funkcí 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 tečné nadroviny k nadhodnocené konkávní Pengově-Robinsonově stavové rovnici. Další výhoda tohoto kroku je vyhnutí se optimalizaci účelové funkce. Metoda je testována na několika určitých příkladech.

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

Full paper: Martin Jex and Jiří Mikyška, An improved branch and bound algorithm for phase stability testing of multicomponent mixtures, Fluid Phase Equilibria, Volume 566, 2023, 113695, ISSN 0378-3812, `https://doi.org/10.1016/j.fluid.2022.113695`

References

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

[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, 112323, 2020.

[3] Michelsen, M. *The isothermal flash problem. Part I. Stability*. Fluid Phase Equlibria, 9, 1-19, 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 Equlibria, 62, 191-210, 1991.

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

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

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

[8] Nichita D. V., de-Hemptinne J.-C., Gomez S. *Isochoric phase stability testing for hydrocarbon mixtures*. Petroleum Science and Technology, 27, 2177-2191, 2009.

[9] Nichita D. V. *Robustness and efficiency of phase stability testing at VTN and UVN conditions*. Fluid Phase Equilibria, 564, 113624, 2023.

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

[11] Smejkal T., Mikyška J., Kukal J. *Comparison of Modern Heuristics on Solving the Phase Stability Testing Problem.* Discrete and Continuous Dynamical Systems Series S, 14, 1161–1180, 2021.

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

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

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

[15] 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, 50(1): 1367–1395, 2018.

[16] Zhang T., Zhang Y., Katterbauer K., Al Shehri A., Sun S., Hoteit I. *Phase equilibrium in the hydrogen energy chain.* Fuel, 328, 125324, 2022.

[17] Dong X., Liu H., J. Hou, Wu K., Chen Z. *Phase Equilibria of Confined Fluids in Nanopores of Tight and Shale Rocks Considering the Effect of Capillary Pressure and Adsorption Film.* Industrial & Engineering Chemistry Research, 55 (3), 798-811, 2016.

[18] Travalloni L., Castier M., Tavares FW., *Phase equilibrium of fluids confined in porous media from an extended Peng-Robinson equation of state.* Fluid Phase Equilibria, 362, 335-341, 2014.

[19] Zhang T., Li Y., Sun S. *Accelerated Phase Equilibrium Predictions for Subsurface Reservoirs Using Deep Learning Methods.* COMPUTATIONAL SCIENCE - ICCS 2019, Lecture Notes in Computer Science, 623-632, 2019.

[20] Li Y., Zhang T., Sun S. *Acceleration of the NVT Flash Calculation for Multicomponent Mixtures Using Deep Neural Network Models.* Industrial & Engineering Chemistry Research, 58 (27), 12312-12322, 2019.

The Early-Universe and the $S_{q,\delta}$ Entropy*

Jaroslav Kňap
knapjaro@fjfi.cvut.cz

study programme: Mathematical Engineering
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. In recent years various generalizations of entropy have been employed to examine diverse plethora of physical issues from statistical perspective. These include applications of non-extensive entropy in deriving GUPs (and EUPs), which are of interest due to allowing to account for minimal length scale; applications to cosmological problems, such as possibility to obtain accelerated expansion phase with ordinary matter; or description of entropy of black holes in extensive manner. In note this we investigate feasibility of combined entropy, leveraging properties of two different entropy generalizations, specifically we will be employing two-parameter entropic functional $S_{q,\delta}$ introduced by Tsallis [1]. We will attempt to apply this entropic functional to the case of black holes. We end the text by discussing relation of Tsallis entropy to conformal and scale symmetry.

Keywords: black holes, non-extensive entropy, scale symmetry

Abstrakt. V poslednich letech byly různe generalizace entropie použity při zkoumání různorodých fyzikálních problemů ze statistické perspektivy. Tyto zahrnují aplikace ne-extenzivních entropií při odvozování GUP (a EUP), které jsou zajímavé jelikož umožnují zahrnout efekt minimalní délkové škály; aplikace na kosmologické problémy, jako například možnost akcelerované expanze vesmíru pouze s běžnou hmotou; nebo popis entropie černých děr v extenzivní formě. V tomto článku prozkoumáme užitečnost kombinované entropie, která využívá vlastnosti dvou různých zobecnění entropie, konkrétně použijeme dvou parametrický entropický funkcionál $S_{q,\delta}$ ktery byl poprvé popsán Tsallisem v [1]. My se pokusíme tento entropický funkcionál aplikovat na příklad černých děr. Text je zakončen diskuzí o vztahu Tsallisovy entropie ke konformní a škálové symetrii.

Klíčová slova: černé díry, ne-extenzivní entropie, škálová symetrie

1 Introduction

It was noted already by Boltzmann that the Boltzmann–Gibbs (BG) statistics is not suitable for description of systems exhibiting long-range interactions (such as gravitational systems) nor for systems with strong correlations (of either classical or quantum nature). [2]. The reason for this, in the case of system with long-range attractive interactions, is that the BG statistics assumes the partition function to have finite value, which is not the fulfilled in such systems, as the energy can decrease without limit violating the

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

assumption of finiteness. In systems with strong correlations assumption of additivity of entropy is violated, as e.g. entanglement entropy present large contribution to total entropy. It is thus likely that BG entropy is inadequate for proper description of the early Universe, such as inflationary Universe, and we should be searching for more fitting alternatives. We will specifically focus on the issue of proper entropic functional for such systems, as black hole entropy (and more generally horizon entropy) provides us with a possible guide.

Over the last century, in the search for a more general framework for statistical description of physical systems numerous proposals for non-BG entropies were put forward. To name a few Renyi entropy, entropies derived from superstatistics or spectral statistics, and also Tsallis S_q and S_δ entropies [2],[1]. It is these last ones that are of primary interest to us. In the rest of this note, we will focus on Tsallis $S_q \equiv S_{q,1}$ and $S_\delta \equiv S_{1,\delta}$ entropies, and their the two-parametric generalization — entropic functional $S_{q,\delta}$. We will start with quick overview of their characteristics, before moving on to application of the more general $S_{q,\delta}$ to early-universe cosmology and links to conformal symmetry.

2 Tsallis S_q entropy

The Tsallis entropy S_q was first put forward by Tsallis in 1988 [3] as a candidate for generalization of BG entropy, and lead to formulation of non-extensive $q-$generalized statistics. Since then the S_q entropy (and q-generalized statistics) has found numerous applications in various branches of physics, including for example turbulence in plasma, particle production in QCD processes [4], non-linear QM, generalized uncertainty principles (GUPs) [5], and many others [2]. Of special attention should be the link to QCD processes, as it can be considered a hint that at high energies it would be relevant for gravity as well due to postulated links between gravity and gauge theories, such as gravity/gauge duality [6] or 'gravity = gauge × gauge' conjecture [7].

S_q has the following prescription

$$S_q = k_b \frac{\sum_i p_i^q - 1}{1-q} = k_B \sum_{i=1}^{W} p_i \left(\ln_q \frac{1}{p_i} \right) \tag{1}$$

where in the second expression $\ln_q$ is the $q-$logarithm defined as

$$\ln_q (x) = \frac{x^{1-q} - 1}{1-q}. \tag{2}$$

It can be easily checked that Tsallis entropy reduces to BG entropy in the limit $q \to 1$.

Tsallis entropy maintains most properties of BG entropy, being non-negative, expansive , however it is non-additive, with the following rule for addition

$$S_q (A+B) = S_q (A) + S_q (B) + (1-q) S_q (A) S_q (B). \tag{3}$$

From the above it is clear S_q can be either sub-, or supra-additive depending on the value of entropic index q. We stress that name "non-extensive entropy/statistics" is a bit of a misnomer, since typically a system does have a single appropriate value of q (which

can be different from 1) for which S_q is extensive quantity. More appropriate would be appellation 'non-additive' entropy/statistics since the entropy actually is non-additive. However, by now the term "non-extensive statistics" has entered common parlance, so we are using it as well.

The entropic index q has effect on probabilities, specifically for $q < 1$ it results in effective enhancement of rate of rare events, whereas $q > 1$ enhances frequent events [2]. From the physical perspective q can be considered as carrier of information about intrinsic fluctuations of the physical system, and it can be shown it is in general suitable in the situations where the described objects have some additional structure [8]. Effectively then, the parameter q is determined a priory from the microscopic dynamics of the system studied, however in practice it is treated as fitting parameter, since those dynamics are typically unknown.

The non-additive Tsallis q-entropy is an appropriate entropy for systems with quasi-power scaling of quantities, i.e. where the number of states scales with the number of system's elements N as

$$W(N) \approx BN^{\tau}. \tag{4}$$

which are typically systems with strong correlations.

For this reason we can presume that S_q entropy should be relevant in the early universe prior to decoupling of various degrees of freedom, e.g. during reheating inflaton field and SM fields were strongly coupled and correlated. Similarly, the SM fields had strong correlations among each other in early post-inflationary universe, fact used in interpretation of CMB and its relation to cosmic structure formation. It is then natural to apply entropic functional which inherently accounts for these correlations.

Of additional interest to us is that this entropy also comes into play when deriving GUP relations [5]. Emergence of a generalized uncertainty principle (GUP) is a generic prediction of theories of quantum gravity, these then imply a minimal length of the order of the Planck length. This minimal length then affects the phase space structure by modifying the elementary cell volume, which becomes momentum-dependent. It can then be shown that statistics maximizing the entropy are non-Gaussian, and is significant for high energies. Additionally, these non-Gaussian statistics produce the same effects as the GUPs, and so describe the same underlying physics from the different viewpoints. These non-Gaussian statistics can be shown to be of the same type as q-generalized Tsallis statistics [9].

Finally, there is an interesting relation to Jackson derivate (which is derivative employing dilation operator instead of translation operator), where

$$S_q = D_q \sum_i p_i^x. \tag{5}$$

In this sense, Tsallis entropic index q is related to how the behaviour differs on scales separated by q. This link to scaling and its (a)-symmetry, should not be surprising, as the original development of S_q entropy was inspired by multifractal behaviour. Fractals (and by extension multifractals) need to possess scale symmetry only asymptotically for vanishing length scales. However, for fractals/multifractals generated by iterative processes, scale symmetry holds for a finite discrete range of scales, these are called self-similar fractals/multifractals.

3 Tsallis S_δ entropy

In 1972 A. Bekenstein first postulated that black holes possess entropy, and proposed it should be proportional to its surface area [10]. The constant of proportionality was derived 3 years later by Hawking [11], fully establishing Bekenstein-Hawking entropy S_{BH}. Feature of this entropy that immediately captured interest of physicists is the fact that entropy S_{BH} is proportional to area A of the surface of the black hole, not to its volume. This has lead to in depth research into holography in the context of gravitational physics, and indeed to dualities such as AdS/CFT [12]. More generally, in d-dimensional strongly correlated quantum systems exists an area law for entropy, i.e. $S_{BG} \approx L^{d-1} = L^{\alpha}$. Looked at from another angle, the scaling factor α determines dimensionality of the problem, and then $\alpha + 1$ can be considered to be the fractal dimension of the problem.

The different scaling has lead to research into a second class of non-additive entropies, the so-called S_δ entropy [1]. In contrast with S_q entropy, S_δ entropy is relatively recent addition, motivated by insufficiency of other entropy generalizations to provide extensive entropy for black holes.

The formulae defining the S_δ is as follows

$$S_\delta \;=\; k_B \sum_{i=1}^{W} p_i \left(\ln \frac{1}{p_i} \right)^{\delta} , \tag{6}$$

which for, e.g. black holes is extensive quantity [1] for appropriate values of δ, and is in general of interest when considering systems with sub-exponential (but non-polynomial) scaling of quantities, i.e. where the number of states scales with the number of system's elements N as

$$W(N) \;\approx\; C\eta^{N^{\gamma}}; \quad C > 0;\, \eta > 1; 0 < \gamma < 1\,. \tag{7}$$

The practical outcome of the δ parameter can be summarized as rescaling of the entropy, e.g. $S^{\delta}_{BG} = S_\delta$, much more straightforward relation than in the case of S_q entropy. For equal probabilities we can write for the entropy S_δ

$$S_\delta = k_B \ln^{\delta} W. \tag{8}$$

Application to black holes is straightforward. We notice that for black holes we have behaviour of type $\ln W(L) \propto L^{d-1}$, i.e. the same scaling in 7. This than implies that black hole is suitable for description by S_δ entropy, being extensive for $\delta = d/(d-1)$. For equal probabilities we easily obtain $S_{\delta=d/(d-1)} \propto (S_{BG})^{d/(d-1)}$.

For the reason above, S_δ can be considered as the proper extensive entropy for BHs, and by extension for gravitating systems with (event or causal) horizons. Recently it has been used to derive modified cosmology (Tsallis cosmology) where for appropriate values of the parameter δ it was possible to obtain the correct accelerated phase of the Universe even with ordinary matter [13]. Regarding physical interpretation of the δ parameter, similar entropy (called Barrow entropy) for black holes was proposed by Barrow in [14], where $\Delta = 2(\delta - 1)$ is parameter describing fractal structure of the surface of black hole due to quantum deformations.

4 Two-parameter entropic functional $S_{q,\delta}$

In the previous sections we have discussed the distinct entropies, S_q and S_δ. We can notice that both have features which are desirable for entropy in quantum cosmology. S_q has link to quadratic GUPs, which are expected to play role at sufficiently high energies as they are generic prediction of different approach to quantum gravity, such as: string theory, loop string gravity, doubly special (and de Sitter) relativity, etc, all of which imply finite minimal length scale. S_δ on the other hand provides possibility of describing black hole (and horizon) entropy in terms of extensive quantity, which can be a desirable feature.

Additionally, there is evidence that high-energy scattering processes are better described by q-statistics (along with S_q entropy) [4]. There are also conjectures on the relation of gravitational and gauge theories, such as gauge-gravity duality originating from string theory [6], or "gravity = gauge × gauge" relating scattering amplitudes [7]. These two facts combined suggest a link between S_q entropy and by extension q-statistics and theories of gravity at high energies, at least on level of scattering amplitudes.

As a result of the above considerations, an appropriate merger of both kinds of aforementioned entropies would be a natural step in a statistical/thermodynamic description of the early Universe. In Ref. [2] a natural two-parameter merger of S_q and S_δ was introduced. These, so called $S_{q,\delta}$ entropies are defined as

$$S_{q,\delta} \;=\; k_B \sum_{i=1}^{W} p_i \left(\ln_q \frac{1}{p_i}\right)^{\delta} , \tag{9}$$

which can be extensive for systems with both sub-exponential scaling and also polynomial scaling, i.e. where the number of internal configurations scales with the number of system's elements N as

$$W(N) \;\approx\; D N^{\tau} \eta^{N^{\gamma}} . \tag{10}$$

We can take the perspective that this allows us to take into account sub-leading corrections to the entropy. For equal probabilities the entropic functional $S_{q,\delta}$ can be written as

$$S_{q,\delta} = k_B \left(\ln_q W\right)^{\delta} , \tag{11}$$

fact we will soon leverage.

This suggestion that $S_{q,\delta}$ could possibly take into account sub-leading correction is interesting from perspective of black hole entropy, since it is well known that universal leading-order correction to the Bekenstein-Hawking entropy is a logarithmic term [15], resulting in entropy of the form

$$S_{BH-corrected} \;\approx\; S_{BH} + \alpha \ln S_{BH} + \kappa , \tag{12}$$

where α and κ depend on details of the theory.

The presence of logarithmic correction is closely related to conformal anomaly [16]. Due to its nature as universal correction, which can be derived from many different considerations (Euclidean action, conformal anomaly, GUPs, etc.), it is natural to consider (12) as an appropriate extension of Bekenstein-Hawking entropy (at least in the functional

form). We are then interested if we can obtain such form of entropy from assumption that $S_{q,\delta}$ is suitable for description of black holes.

We will be starting from assumption that the entropy functional $S_{q,\delta}$ describes (with some degree of precision) entropy of black hole, then starting from expressions for scaling of number of internal configurations and for entropy $S_{q,\delta}$ in case of equal probabilities

$$\begin{aligned} W\left(L^d\right) &\approx D\left(L^d\right)^{\tau}\eta^{\left(L^d\right)^{\gamma}}, \\ S &= k_B\left(\ln_q W\right)^{\delta}, \end{aligned} \tag{13}$$

we obtain

$$S_{BH-q,\delta} = k_b\left(\frac{W^{1-q}-1}{1-q}\right)^{\delta}, \tag{14}$$

by combination of the two. Additionally, we demand that the total number of internal configurations W is such that $\ln W = S_{BH}$, so that we may recover the usual Bekenstein–Hawking formula in the Boltzmann limit $q \to 1$ and $\delta \to 1$.

We expand the formula (14) around $q = 1$ as we are interested in small deviations from extensivity. The reasoning is that cosmological observations and other estimates from high energy physics show that if S_q entropy is used the entropic index q is typically around the value 1.218 [17] [18]. Later on this assumption of small deviation will be justified by consistency of the result with the assumption, up to some finite energy scale. This limitation to certain energy scales should not be suprising, as QFT effects typically exhibit scale dependency.

After short algebraic operations, we obtain

$$S_{BH-q,\delta} \approx k_b\left(\ln W - \frac{1}{2}\ln^2 W\,(q-1)\right)^{\delta}, \tag{15}$$

where we neglect terms of order $(q-1)^2$ and greater.

We can express $\ln W$ in terms of the L^d by using the scaling relation for number of configurations. Taking its logarithm we obtain

$$\ln W \approx EL^{d/\delta} + \ln L^{d\tau} \tag{16}$$

where $E = \ln\eta$. Setting

$$\begin{aligned} q &= 1 + 2\left(\ln L^{d\tau} - \ln L^{d/\delta}\right)\left(EL^{d/\delta} + \ln L^{d\tau}\right)^{-2} \\ \delta &= d/\left(d-1\right) \end{aligned} \tag{17}$$

we obtain the usual form of the black hole entropy along with logarithmic corrections

$$S_{BH-q,\delta} \approx k_B\left(EL^{d-1} + \ln L^{d-1}\right)^{d/(d-1)}. \tag{18}$$

Of course this is only approximate calculation, full calculation would be far more complex as it would require proper treatment of $\ln_q$ function, which does not share some of the properties of logarithmic function we have leveraged.

Let us draw attention to the fact that parameter q is running with the energy scale (in this case specifically with the mass of BH). This should not be surprising, as the entropy is a measure of the physical degrees of systems, which are scale-dependent in QFT. Hence, the outcome of the calculation is rooted in quantum field theory considerations, which although not taken into account by Tsallis in his original formulation of non-extensive thermodynamics, are in principle mandatory when one tries to extend Tsallis entropy to a more general QFT framework. Additionally, since the correction we calculate is purely quantum in nature, appearance of QFT behaviour such as running should not be surprising. So, purely from the assumption on the scaling behaviour of the degrees of freedom of BH we were able to (approximately) derive black hole entropy along with logarithmic corrections.

We note that it was the modification arising from the q deformation that lead to presence of the logarithmic correction, showcasing the link to GUPs as they can be used to derive the same term. The modification from δ plays comparatively small role, only leading to rescaling of the dimension of entropy so that is extensive quantity.

Black hole horizons are not the only horizons present in general relativity (or its generalizations). Cosmological horizons are also a type of event horizon, and so can be associated with entropy along the same lines as black holes. This fact was firstly recognized by Gibbons, hence Gibbons–Hawking entropies [19]. From the successful application of $S_{q,\delta}$ entropy to the case of black hole horizon we can conclude that $S_{q,\delta}$ entropy should also be relevant to the cosmological horizons, and hence large-scale cosmology.

5 Summary and conclusion

We applied the non-extensive two-parameter entropic functional $S_{q,\delta}$ to the case of primordial black hole. From this starting assumption (along with assumption on scaling of number of microstates), we have derived the Bekenstein-Hawking entropy along with logarithmic correction arising from quantum corrections.

Additionally, by considering relation of Tsallis S_q entropy to GUPs and S_δ entropy to gravitational horizons, it is reasonable to assume that two parametric entropy functional $S_{q,\delta}$ should play a significant role in the early-Universe cosmology. Since black hole horizons are only one specific type of horizon present in theories of gravity, we further extrapolate that the $S_{q,\delta}$ entropy is suitable for description of cosmological horizons as well. For this reason, we should expect it to play role in cosmology of early universe, where both borizon and GUPs effects can play role.

Regarding extremely early Universe, we stress that one would expect GUPs that have also higher-order corrections beyond quadratic ones, necessitating entropy functionals beyond S_q and $S_{q,\delta}$. These would rise to relevancy as the energy scale increases. Interestingly, studies of cosmology resulting from thermodynamics considerations with varying non-extensive parameter has been recently addressed in the context of modified cosmological models [20]. There, it has been shown that such cosmologies naturally lead to inflationary de Sitter solutions, suggesting certain correspondence between cosmology derived from power-law $f(R)$ gravity and non-extensive cosmology. The correspondence can be used to relate power-law exponent and the non-extensive exponent, at least within de Sitter geometry.

We end this by returning to the link to scaling symmetry. Q-statistics, based on S_q entropy, often appear when the systems have fractal structure (so called 'thermofractal' [21]) in energy-momentum space. As the conformal (and hence scaling) symmetry is spontaneously breaking in inflationary universe, we posit that scale symmetry could leave an imprint after spontaneous symmetry breaking in the form of fractal/multifractal structures. Indeed, recent work in this direction does suggest that scale symmetry breaking can result in such features, such as fractal energy spectrum [22]. It is to be remarked that the definition of multifractal, as well as the simpler definition of fractal set, are formulated in terms of limits for some vanishing length scale. In other words, the scale symmetry only needs to hold asymptotically for vanishing length scales. There is a class of multifractals in which scale symmetry holds in a finite range of scales, namely the self-similar fractals or multifractals. Therefore we argue that after SSB of conformal symmetry, such fractal or multifractal behaviour could be present, further justifying application of q-statistics and S_q entropy as a consequence of conformal symmetry.

References

[1] C. Tsallis and L. J. L. Cirto, Eur. Phys. J. C **73**, 2487 (2013).

[2] C. Tsallis, *Introduction to Non-Extensive Statistical Mechanics: Approaching a Complex World*, (Springer, Berlin, 2009).

[3] C. Tsallis: *Possible generalization of Boltzmann-Gibbs statistics*, (Journal of Statistical Physics. 52 (1–2): 479–487. , 1988)

[4] G. Wilk, Z. Włodarczyk: *Some Non-Obvious Consequences of Non-Extensiveness of Entropy*, (Entropy. 2023; 25(3):474) https://doi.org/10.3390/e25030474

[5] P. Jizba, J.A. Dunningham and M. Prokš, Entropy **23**, 334 (2021).

[6] M. Ammon and J. Erdmenger: *Gauge/gravity duality: Foundations and applications*, (Cambridge University Press, Cambridge, 2015)

[7] Z. Bern, J. J. M. Carrasco, H. Johansson: *Perturbative Quantum Gravity as a Double Copy of Gauge Theory*, (Phys. Rev. Lett. 105:061602, 2010)

[8] G. Wilk, Z. Włodarczyk: *Example of a possible interpretation of Tsallis entropy*, (Physica A: Statistical Mechanics and its Applications, Vol. 387, 19, p. 4809-4813.)

[9] H. Shababi, K. Ourabah: *Non-Gaussian statistics from the generalized uncertainty principle*, (Eur. Phys. J. Plus 135, 697, 2020)

[10] A. Bekenstein: *Black holes and the second law*, (Lettere al Nuovo Cimento. 4 (15): 99–104. , 1972)

[11] S. W. Hawking: *Particle creation by black holes*, (Communications in Mathematical Physics. 43 (3): 199–220. 1975)

[12] J. Maldacena: *The Large N limit of superconformal field theories and supergravity*, (Advances in Theoretical and Mathematical Physics. 2 (4), 1998)

[13] P. Jizba, G. Lambiase, Eur. Phys. J. C **82**, 1123 (2022).

[14] J.D. Barrow Phys. Lett. B, 808 (2020).

[15] D. V. Fursaev, Temperature and entropy of a quantum black hole and conformal anomaly: (Phys. Rev. D51 (1995) 5352–5355) [hep-th/9412161]

[16] R. G. Cai, L. M. Cao and N. Ohta: *Black Holes in Gravity with Conformal Anomaly and Logarithmic Term in Black Hole Entropy*, (JHEP 04, 082 (2010)) [arXiv:0911.4379 [hep-th]]

[17] C. Beck: *Superstatistics in high-energy physics*, (Eur. Phys. J. A 40, 267, 2009)

[18] T.S. Biro, V.G. Czinner: *A q-parameter bound for particle spectra based on black hole thermodynamics with Rényi entropy*, (Phys. Lett. B 726, 861, 2013)

[19] G.W. Gibbons, Phys. Rev. D, **15**, 2738 (1977).

[20] S. Nojiri, S.D. Odintsov, E.N. Saridakis: *Modified cosmology from extended entropy with varying exponent*, (Eur. Phys. J. C 79, 242, 2019)

[21] A. Deppman, J. A. S. Lima: *Thermofractals and the Nonextensive Finite Ideal Gas*, (Physics 2021, 3(2), 290-301)

[22] O. Ovdat, J. Mao, Y. Jiang, E. Y. Andrei, E. Akkermans: *Observing a scale anomaly and a universal quantum phase transition in graphene*, (Nature Communications Vol. 8, Article number: 507 (2017))

CNN Ensemble Robust to Rotation Using Radon Transform

Václav Košík
kosikvac@fjfi.cvut.cz

study programme: Mathematical Engineering
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. A great deal of attention has been paid to alternative techniques to data augmentation in the literature. Their goal is to make convolutional neural networks (CNNs) invariant or at least robust to various transformations. In this paper, we present an ensemble model combining a classic CNN with an invariant CNN where both were trained without any augmentation. The goal is to preserve the performance of the classic CNN on nondeformed images (where it is supposed to classify more accurately) and the performance of the invariant CNN on deformed images (where it is the other way around). The combination is controlled by another network which outputs a coefficient that determines the fusion rule of the two networks. The auxiliary network is trained to output the coefficient depending on the intensity of the image deformation. In the experiments, we focus on rotation as a simple and most frequently studied case of transformation. In addition, we present a network invariant to rotation that is fed with the Radon transform of the input images. The performance of this network is tested on rotated MNIST and is further used in the ensemble whose performance is demonstrated on the CIFAR-10 dataset.

Keywords: CNN, rotation invariance, equivariance, Radon transform, network ensemble

Abstrakt. Alternativním technikám k datové augmentaci bylo v literatuře věnováno mnoho pozornosti. Jejich cílem je učinit konvoluční neuronové sítě invariantní nebo alespoň robustní vůči různým obrázkovým transformacím. V tomto článku představujeme ensemble kombinující klasickou konvoluční síť s invariantní konvoluční sítí, přičemž obě jsou trénované bez jakékoliv augmentace. Cílem je dosáhnout přesnosti klasické sítě na nedeformovaných obrázcích, na kterých předpokládáme, že bude úspěšnější, a přesnosti invariantní sítě na deformovaných obrázcích, kde předpokládáme opak. Přesná podoba kombinace se opírá o další sítí, která počítá koeficient, jenž určuje, která ze dvou sítí bude mít při predikci konkrétního vstupu dominantní roli. Tato pomocná síť je také trénovaná a počítaný koeficient na výstupu závisí na míře obrázkové deformace. V experimentech se zaměřujeme na konkrétní příklad deformace, a to rotaci, z důvodu její jednoduchosti a vysoké míry výskytu v literatuře. Dále představujeme konvoluční neuronovou síť invariantní na rotaci. Její konstrukce je založena především na Radonově transformaci vstupních obrázků. Její efektivitu testujeme na datasetu "rotated MNIST" a dále ji využíváme ve zmíněném ensembleu, jehož úspěšnost demonstrujeme na datasetu CIFAR-10.

Klíčová slova: konvoluční neuronové sítě, rotační invariance, ekvivariance, Radonova transformace, ensemble

Full paper: V. Košík, T. Karella, J. Flusser, *CNN Ensemble Robust to Rotation Using Radon Transform*, accepted at International Conference "Image Processing Theory, Tools and Applications (IPTA2023)".

A GENERIC Theory of the Van der Waals Fluid*

Juraj Kováč
kovacjur@fjfi.cvut.cz

study programme: Mathematical Engineering
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. Constructing a thermodynamically consistent dynamical van der Waals theory that is compliant with experimental observations could prove to be a major leap towards a kinetic theory of phase transitions. Here we propose two candidates for such a theory in the so-called GENERIC (general equation for non-equilibrium reversible irreversible coupling) form that safeguards their consistency with the fundamental laws of thermodynamics. Some of the major physical properties of the theories are discussed and the corresponding hydrodynamic equations are derived. Directions of future research are indicated.

Keywords: van der Waals theory, kinetic equation, GENERIC, phase transitions

Abstrakt. Vybudovanie dynamickej van der Waalsovej teórie konzistentnej so zákonmi termodynamiky i experimentálnymi pozorovaniami by nás mohlo výrazne posunúť smerom k ucelenej kinetickej teórii fázových prechodov. Tento text predkladá dvoch kandidátov na takúto teóriu v takzvanom GENERIC tvare, ktorý zaručuje súlad s fundamentálnymi zákonmi termodynamiky. Ponúkame diskusiu vybraných kľúčových fyzikálnych aspektov týchto teórií, ako aj odvodenie zodpovedajúcich hydrodynamických rovníc a naznačujeme smery ďalšieho výskumu.

Kľúčové slová: van der Waalsov plyn, kinetická rovnica, GENERIC, fázové prechody

Introduction

The van der Waals (vdW) theory represents probably the simplest known model that allows for multiphase states[13] and as such is a natural candidate for modelling phase transitions. As a result, establishing a dynamical vdW theory might be a vital step in achieving the elusive goal of a dynamical theory of phase transitions. On the following pages, we propose an alternative construction of the kinetic theory of vdW fluid to the commonly deployed Enskog-Vlasov (EV) equation. The key features that distinguish this theory from ideal gas are finite size of particles in the Enskog collision term[4] and a long-range intermolecular interaction introduced by means of the Vlasov term[14] that causes particle trajectories to divert from the straight lines of the Boltzmann gas. As it turns out, these two features of the vdW gas allow for the phenomenon of phase transitions[6]

*The author would like to express his sincere gratitude to professor Miroslav Grmela for his hospitality and our fruitful collaboration. This work has also been supported by the Czech Grant Agency, project number 20-22092S.

not present in the Boltzmann theory, both gas-liquid[13, 3] and liquid-solid[2]. Our goal is to construct a theory that provides a description of the kinetic phenomena near the interface and has a straightforward relationship to the equilibrium theory as presented in references[13] or [11], while bypassing the consistency issues of the EV equation with the laws of thermodynamics[7, 1, 12]. This is achieved via a kinetic equation that has the GENERIC (general equation for non-equilibrium reversible irreversible coupling)[8, 9] form. We offer here two such equations, discuss their possible advantages and disadvantages and sketch directions of further analysis to either show their compliance with other research in this domain, whether experimental, numerical or theoretical, or to refute them as insufficient or inaccurate. These may include providing a description of the corresponding equilibrium theory including an equation of state, analysing corresponding mesoscopic (hydrodynamic) theories, formulating predictions with respect to the structure of the interface layer etc.

1 Two GENERIC van der Waals theories

We are going to propose and subject to inspection two alternative models for the van der Waals gas, outlining prospective directions of further progress and discussing possible advantages and disadvantages of the respective models. The broader framework of our considerations will be the GENERIC (general equation for non-equilibrium reversible-irreversible coupling)

$$\dot{x} = \mathcal{L}(x)E_x + \Xi_{x^*}(x, x^*)|_{x^*=S_x}, \tag{1}$$

where x stands for the state variable, $\mathcal{L}(x)$ is the Poisson bivector, $\Xi(x, x^*)$ is the dissipation potential, $E(x)$ is the energy, and $S(x)$ is the entropy. That is to say that our starting point is the more general, underlying model (1) and proposing a specific model for the van der Waals gas is equal to choosing the state variables x as well as specifying the quantities

$$(\mathcal{L}, E, S, \Xi) \tag{2}$$

of the model, the *constitutive relations*. As we shall see, the clear distinction between the reversible and irreversible part of the dynamics provided by (1) is highly advantageous for imposing (or studying the validity of) the requirements of fundamental physics on a dynamical model.

As already noted, unlike the Boltzmann equation, the EV equation is not of the form (1). Let us first demonstrate the choice of constitutive relations (2) corresponding to the Boltzmann equation itself. We assume an energy functional $E(f)$ given in terms of the 1-particle distribution function f and define the 1-particle energy to be the functional derivative $E^{(p)} := \frac{\delta E}{\delta f} \equiv E_f$. Denoting $\boldsymbol{r}$ and $\boldsymbol{v}$ the position and momentum, respectively, the Boltzmann equation can then be written in the form

$$\frac{\partial f(\boldsymbol{r}, \boldsymbol{v})}{\partial t} = -\frac{\partial \left(f \frac{v_i}{m}\right)}{\partial r_i} + \mathcal{B}(f(\boldsymbol{r}, \boldsymbol{v})) \tag{3}$$

where $\mathcal{B}$ is the Boltzmann collision term. We rewrite this equation in a more abstract form as

$$\frac{\partial f(\boldsymbol{r}, \boldsymbol{v})}{\partial t} = -\frac{\partial (f E^{(p)}_{v_i})}{\partial r_i} + \frac{\partial (f E^{(p)}_{r_i})}{\partial v_i} + \mathcal{B}(f(\boldsymbol{r}, \boldsymbol{v})), \tag{4}$$

where we included the zero term $E^{(p)}_{r_i}$ for later use, the subscript meaning partial derivative. We will now demonstrate that the equation (3) is indeed of the form (1).

Let us start by making the observation that the (more general) Waldmann collision term can be expressed as the functional derivative $\frac{\delta\Xi}{\delta f^*}$ of the mass-action law dissipation potential

$$\Xi^{(W)}(f,f^*) = \int d\boldsymbol{v} \int d\boldsymbol{v}_1 \int d\boldsymbol{v}' \int d\boldsymbol{v}'_1 W(\boldsymbol{v},\boldsymbol{v}_1,\boldsymbol{v}',\boldsymbol{v}'_1,f)\left(e^X + e^{-X} - 2\right) \tag{5}$$

with respect to the conjugate distribution function $f^*(\boldsymbol{r},\boldsymbol{v}) = \frac{\delta S}{\delta f(\boldsymbol{r},\boldsymbol{v})} \equiv S_{f(\boldsymbol{r},\boldsymbol{v})}$, where the thermodynamic force X defined by

$$X = \frac{1}{k_B}\left(f^*(\boldsymbol{r},\boldsymbol{v}) + f^*(\boldsymbol{r},\boldsymbol{v}_1) - f^*(\boldsymbol{r},\boldsymbol{v}') - f^*(\boldsymbol{r},\boldsymbol{v}'_1)\right) \tag{6}$$

was introduced. Setting now S to be the Boltzmann entropy

$$S(f) = -k_B \int d\boldsymbol{r} \int d\boldsymbol{v} f(\boldsymbol{r},\boldsymbol{v}) \ln f(\boldsymbol{r},\boldsymbol{v}), \tag{7}$$

the collision term becomes[11]

$$\mathcal{B}^{(W)}(f) = \Xi^{(W)}_{f^*} = \int d\boldsymbol{v}_1 \int d\boldsymbol{v}' \int d\boldsymbol{v}'_1 \frac{2W(\boldsymbol{v},\boldsymbol{v}_1,\boldsymbol{v}',\boldsymbol{v}'_1,f)}{k_B\sqrt{ff_1f'f'_1}}\left(f'f'_1 - ff_1\right), \tag{8}$$

where $f = f(\boldsymbol{v}), f_1 = f(\boldsymbol{v}_1), f' = f(\boldsymbol{v}'), f'_1 = f(\boldsymbol{v}'_1)$ and the common positional argument $\boldsymbol{r}$ is omitted. Given the corresponding choice of the integral kernel W[15], (8) becomes exactly the Boltzmann collision term $\mathcal{B}(f)$.

Following the choice of the dissipation potential Ξ and the entropy S, which specify the dissipative part of the dynamics, the remaining task is to set the Poisson bivector $\mathcal{L}$ and the energy E such that they yield the Hamiltonian part of the dynamics of equation (3). As the Boltzmann (ideal) gas consists of free, colliding particles, its energy is obviously

$$E(f) = \int d\boldsymbol{r} \int d\boldsymbol{v} f(\boldsymbol{r},\boldsymbol{v}) \frac{\boldsymbol{v}^2}{2m}, \tag{9}$$

which sets the 1-particle energy to be $E^{(p)} = \frac{\boldsymbol{v}^2}{2m}$. Finally, the Poisson bracket corresponding to (3) (which again carries Boltzmann's name), is

$$\{A,B\} = \int d\boldsymbol{r} \int d\boldsymbol{v} f\left(\frac{\partial A_f}{\partial r_i}\frac{\partial B_f}{\partial v_i} - \frac{\partial B_f}{\partial r_i}\frac{\partial A_f}{\partial v_i}\right). \tag{10}$$

Let us recall that if a system is described by the one particle distribution function f as the sole state variable, the reversible part of the evolution equation (1) is given by

$$\frac{\partial f(\boldsymbol{r},\boldsymbol{v})}{\partial t} = \mathcal{L}E_{f(\boldsymbol{r},\boldsymbol{v})}. \tag{11}$$

Here, the energy E plays the role of a generating functional and the operator $\mathcal{L}(f)$, the *Poisson bivector*, follows from the Poisson bracket by $\{A,B\} = \langle A_f, \mathcal{L}B_f\rangle$. Rewriting (10) as

$$\{A,B\} = \int d\boldsymbol{r} \int d\boldsymbol{v} A_f\left(\frac{\partial B_f}{\partial r_i}\frac{\partial f}{\partial v_i} - \frac{\partial B_f}{\partial v_i}\frac{\partial f}{\partial r_i}\right),$$

it can be seen that the Boltzmann Poisson bracket (10) generates an equation of the abstract form (4). Clearly, with the energy (9), the Boltzmann equation (3) is recovered.

Before proceeding with alternative theories of the van der Waals gas, let us summarize some of the key properties and assumptions that render systems of the form (1) consistent with the laws of thermodynamics. We start with the Poisson bracket, which captures the reversible (Hamiltonian) part of the evolution of the system in question and as such is assumed to satisfy (besides antisymmetry and the Jacobi identity) the degeneracy conditions

$$\mathcal{L}S_x = 0, \qquad \mathcal{L}N_x = 0 \tag{12}$$

In other words, both the entropy S and the total number of particles $N = \int d\boldsymbol{r} \int d\boldsymbol{v} f$ are assumed to be *Casimirs* of the Poisson bracket. Evidently, this implies that the reversible part of the dynamics preserves both. The Poisson bracket is taken to be the Boltzmann Poisson bracket (10). It is a back-of-the-envelope calculcation to verify that any entropy functional of the form

$$S = \int d\boldsymbol{r} \int d\boldsymbol{v} \nu(f), \tag{13}$$

where $\nu : \mathbb{R} \to \mathbb{R}$ is sufficiently smooth (e.g. of class $\mathcal{C}^2([0,1])$), is a Casimir of the Boltzmann Poisson bracket (10). Let us also note that under the assumptions (12), equation (11) can be equivalently rewritten

$$\frac{\partial f(\boldsymbol{r},\boldsymbol{v})}{\partial t} = T\mathcal{L}\Phi_{f(\boldsymbol{r},\boldsymbol{v})}, \tag{14}$$

with $\Phi(f,T,\mu) := -S(f) + \frac{1}{T}E(f) - \frac{\mu}{T}N(f)$ denoting the thermodynamic potential. We are going to 'abuse' this observation later in the text by introducing a kinetic equation of the form (14) with an entropy that is not a Casimir of the Boltzmann Poisson bracket (10).

The irreversible part of the dynamics is expressed via the conjugate variables $x^* = S_x$ as well as the dissipation potential Ξ. First of all, let us specify that by *irreversibility* we mean dynamics that is even with respect to the time reversal transformation (TRT) $(\boldsymbol{r} \to \boldsymbol{r}, \boldsymbol{v} \to -\boldsymbol{v}, t \to -t)$, i.e.

$$\Xi(x, x^*) \xrightarrow{\text{TRT}} \Xi(x, x^*).$$

The second law of thermodynamics is encapsulated into the requirement

$$\langle \Xi_{x^*}, x^* \rangle \geq 0$$

($\langle \cdot, \cdot \rangle$ again denotes scalar product), where the left-hand side becomes the temporal derivative of the entropy $\frac{dS}{dt}$ if the conjugate state variables S_x are substituted for x^*. The equilibrium nature of the state $x^* = 0$ is expressed by $\Xi|_{x^*=0} = 0$ and the conservation of mass and energy is guaranteed by the degeneracy conditions

$$\langle \Xi_{x^*}, M_x \rangle = 0, \qquad \langle \Xi_{x^*}, E_x \rangle = 0.$$

Before proceeding, let us note that the energy of the system will be assumed in the form

$$E(f) = \int d\boldsymbol{r} \int d\boldsymbol{v} f(\boldsymbol{r},\boldsymbol{v}) \left\{ \frac{\boldsymbol{v}^2}{2m} + \frac{1}{2} \int d\boldsymbol{r}_1 \int d\boldsymbol{v}_1 f(\boldsymbol{r}_1,\boldsymbol{v}_1) \phi(|\boldsymbol{r} - \boldsymbol{r}_1|) \right\}. \tag{15}$$

1.1 van der Waals-Vlasov equation I

The first dynamic van der Waals theory we will investigate is obtained by replacing the entropy (7) with the van-der-Waals-like entropy

$$S(f) = k_B \int d\boldsymbol{r} \int d\boldsymbol{v} \underbrace{f(\boldsymbol{r},\boldsymbol{v}) \left[\ln(h^3 f(\boldsymbol{r},\boldsymbol{v})) - 1 - \ln\left(1 - \frac{b f(\boldsymbol{r},\boldsymbol{v})}{N_A}\right)\right]}_{=:\eta(f(\boldsymbol{r},\boldsymbol{v}))}, \tag{16}$$

where $b > 0$ represents the (duly normalized) volume occupied by N_A molecules of the gas. The additional term $\ln\left(1 - \frac{bf(\boldsymbol{r},\boldsymbol{v})}{N_A}\right)$ obviously accounts for the finite size of the van der Waals particles.

First, note that this entropy is of the form (13) and is hence a Casimir of the Boltzmann Poisson bracket (10). This fact guarantees that the Hamiltonian part (generated by the Boltzmann Poisson bracket) of the ensuing kinetic equation conserves entropy. The collision term, however, will now differ from the Boltzmann collision term (8) due to the alteration of $f^* = S_f$. The idea here is to replace the Enskog collision term with a modification of the Boltzmann collision term. Since this new term will preserve locality of collisions, we expect the effects of dissipation to be somewhat weaker compared to the EV theory. The modified collision term corresponding to the entropy (16) takes the form

$$\begin{aligned}\mathcal{B}^{(vdW)}(f) = &\int d\boldsymbol{v}_1 \int d\boldsymbol{v}' \int d\boldsymbol{v}'_1 \frac{2W(\boldsymbol{v},\boldsymbol{v}_1,\boldsymbol{v}',\boldsymbol{v}'_1,f)}{k_B\sqrt{f f_1 f' f'_1}} \times \\ &\times \left[f' f'_1 - f f_1 + b\,(f f_1 + f' f'_1)(f' + f'_1 - f - f_1)\right] + \mathcal{O}(b^2),\end{aligned}$$

defining the kinetic equation corresponding to this model to be

$$\frac{\partial f}{\partial t} = -\frac{\partial\left(f\frac{v_i}{m}\right)}{\partial r_i} + \frac{1}{m}\frac{\partial \Psi}{\partial r_i}\frac{\partial f}{\partial v_i} + \mathcal{B}^{(vdW)}(f), \tag{17}$$

where

$$\Psi(\boldsymbol{r}) = \int d\boldsymbol{r}' n(\boldsymbol{r}')\phi(|\boldsymbol{r} - \boldsymbol{r}'|), \tag{18}$$

ϕ being the potential. Note that equation (17) is of the abstract form (4). Obviously, the zeroth-order term is the Boltzmann collision term. The physical interpretation of this first-order correction, which we expect to account (partially, at least) for the finite size of the particles, is somewhat unclear.

It is worth noting that modifying the entropy has important consequences for the equilibrium theory, too. The equilibrium distribution function f_{eq} obtained by entropy maximization with respect to the entropy (16), given a knowledge of the energy (15) and the number of particles N, will no longer be Maxwellian. More precisely, denoting $\beta = \frac{1}{2mk_BT}$, we find the equilibrium (maximum entropy) solution to be

$$f_{eq}(\boldsymbol{r},\boldsymbol{v}) = \frac{N_A}{b}\frac{W(x)}{W(x)+1}, \tag{19}$$

where $x(\boldsymbol{r},\boldsymbol{v}) = \frac{b}{N_A h^3}\exp\left(-\beta\boldsymbol{v}^2 - \psi(\boldsymbol{r}) + \frac{\mu}{k_BT}\right), \psi(\boldsymbol{r}) = \frac{1}{k_BT}\Psi(\boldsymbol{r})$, with Ψ given in (18). While this is only an implicit equation for f_{eq} as long as the Vlasov term ψ (which depends

on n and hence implicitly on f) is included, it can be used to infer the (non-)Maxwellian character of the equilibrium. If we now expand (19) into a Taylor series in b, we have

$$f_{eq} = \frac{1}{h^3}\exp\left(-\psi(\boldsymbol{r})\right)\exp(-\beta\boldsymbol{v}^2) - \frac{2b}{h^6 N_A}\exp\left(-2\psi(\boldsymbol{r})\right)\exp(-2\beta\boldsymbol{v}^2) + \mathcal{O}(b^2).$$

Clearly, for $b \neq 0$ we have a non-Maxwellian equilibrium, as the zeroth-order (Maxwellian) term then obtains an additional first-order pseudo-Maxwellian correction.[1]

As kinetic theories are more conveniently tested (whether numerically or experimentally) on more macroscopic levels, let us now proceed with formulation of the (generalized) hydrodynamics that corresponds to the kinetic equation (17). There are, of course, multiple ways to do this, depending on the specific order of approximation as well as on the choice of the state variables. We will demonstrate two of these theories, one in this section and one in the next one. The common feature of both will be that we will only take into account the Hamiltonian part of the dynamics, the Poisson bracket.

In the first example, we take a more general approach. First we define the state variables to be the moments of the distribution function

$$\begin{aligned} \rho(\boldsymbol{r}) &= \int d\boldsymbol{v} f(\boldsymbol{r},\boldsymbol{v}), \qquad u_i(\boldsymbol{r}) = \int d\boldsymbol{v} v_i f(\boldsymbol{r},\boldsymbol{v}), \qquad b_{ij}(\boldsymbol{r}) = \int d\boldsymbol{v} v_i v_j f(\boldsymbol{r},\boldsymbol{v}), \\ s(\boldsymbol{r}) &= \int d\boldsymbol{v}\eta\left(f(\boldsymbol{r},\boldsymbol{v})\right), \quad a_i(\boldsymbol{r}) = \int d\boldsymbol{v} v_i \eta\left(f(\boldsymbol{r},\boldsymbol{v})\right), \end{aligned} \tag{20}$$

with η defined in (16). Here ρ represents the number density, u_i the momentum density, b_{ij} the stress tensor, s the entropy density and a_i can be interpreted as the entropic flux multiplied by the mass density. Note the presence of the entropic moments s and a_i. We are going to admit dependence on these moments for as long as possible in order to obtain their evolution equations.

The next step is to express the Poisson bracket for functionals dependent exclusively on these moments. We choose to start with the Boltzmann Poisson bracket (10) and insert $A_f = A_\rho + A_{u_i}v_i + A_{b_{ij}}v_iv_j + A_s\eta_f + A_{a_i}\eta_f v_i$ (where we omit formal delta functions) to arrive at

$$\begin{aligned} \{A,B\} = \int d\boldsymbol{r}\, \{\, & \rho\left(\partial_i(A_\rho)B_{u_i} - \partial_i(B_\rho)A_{u_i}\right) \\ & + u_i\left(\partial_j(A_\rho)B_{b_{ij}} - \partial_j(B_\rho)A_{b_{ij}}\right) + u_i\left(\partial_j(A_{u_i})B_{u_j} - \partial_j(B_{u_i})A_{u_j}\right) \\ & + b_{ij}\left(\partial_k(A_{u_i})B_{b_{jk}} - \partial_k(B_{u_i})A_{b_{jk}}\right) + b_{ik}\left(\partial_j(A_{u_i})B_{b_{jk}} - \partial_j(B_{u_i})A_{b_{jk}}\right) \\ & + b_{ij}\left(\partial_k(A_{b_{ij}})B_{u_k} - \partial_k(B_{b_{ij}})A_{u_k}\right) + b_{ij}\left(\partial_k(A_\rho)B_{c_{ijk}} - \partial_k(B_\rho)A_{c_{ijk}}\right) \\ & + s\left(\partial_i(A_s)B_{u_i} - \partial_i(B_s)A_{u_i}\right) + s\left(\partial_i(A_\rho)B_{a_i} - \partial_i(B_\rho)A_{a_i}\right) \\ & + a_i\left(\partial_j(A_{a_i})B_{u_j} - \partial_j(B_{a_i})A_{u_j}\right) + a_i\left(\partial_j(A_s)B_{b_{ij}} - \partial_j(B_s)A_{b_{ij}}\right) \\ & + a_j\left(\partial_i(A_{u_j})B_{a_i} - \partial_i(B_{u_j})A_{a_i}\right) + a_j\left(\partial_i(A_s)B_{b_{ij}} - \partial_i(B_s)A_{b_{ij}}\right) \\ & + \int d\boldsymbol{v}\,[\, \eta_f\partial_j(\eta)\left(A_{a_j}B_s - B_{a_j}A_s\right) \\ & + \eta_f\partial_{v_j}(\eta)v_k\left(A_s\partial_j(B_{a_k}) - B_s\partial_j(A_{a_k})\right) \\ & + \eta_f\partial_{v_j}(\eta)v_k\left(A_{a_k}\partial_j(B_s) - B_{a_k}\partial_j(A_s)\right) \\ & + \eta_f v_l B_{a_l}\left(\partial_j(\eta)A_{a_j} - \partial_{v_j}(\eta)v_k\partial_j(A_{a_k})\right)\,]\,\}, \end{aligned} \tag{21}$$

[1] The higher-order terms are also, evidently, pseudo-Maxwellian.

with $\eta_f = \ln(h^3 f) - \ln\left(1 - \frac{bf}{N_A}\right) + \frac{bf}{N_A - bf}$, $\eta_{ff} = \frac{1}{f} + \frac{b}{N_A - bf}$. The last integral can, of course, be rewritten in a *manifestly* antisymmetric form that is, however, more complicated. We therefore prefer to keep it in this form. Note that the corresponding terms for the moments of f are much easier to handle. Here the situation is more complicated due to the presence of the function $\eta(f)$, which, in the moments of f, is replaced with f itself or, in other words, with $\sigma(f)$ given by $\sigma(x) = x$, resulting in $\sigma_f = 1, \partial_j \sigma = \partial_j f$, etc. These terms will not play a role in the equations of motion, however, if we constrain ourselves to energies which only depend on $\rho, \boldsymbol{u}$ and $\boldsymbol{b}$. Then the resulting equations take the form

$$\begin{aligned}
\partial_t s &= -\partial_j \left(s E_{u_j}\right) - 2\partial_j \left(a_k E_{b_{kj}}\right), \\
\partial_t a_i &= -\partial_j \left(a_i E_{u_j}\right) - s\partial_i E_\rho - a_j \partial_i E_{u_j}, \\
\partial_t \rho &= -\partial_j \left(\rho E_{u_j}\right) - 2\partial_j \left(u_k E_{b_{kj}}\right), \\
\partial_t u_i &= -\partial_j \left(u_i E_{u_j}\right) - 2\partial_j \left(b_{ik} E_{b_{jk}}\right) - \rho\partial_i E_\rho - u_j \partial_i E_{u_j} - b_{jk}\partial_i E_{b_{jk}}, \\
\partial_t b_{ik} &= -\partial_j \left(b_{ik} E_{u_j}\right) - u_k \partial_i E_\rho - u_i \partial_k E_\rho - b_{jk}\partial_i E_{u_j} - b_{ij}\partial_k E_{u_j}.
\end{aligned} \tag{22}$$

Probably the most significant shortcoming of this hierarchy is the fact that a large portion of the interrelation between s and $\boldsymbol{a}$ seems to be in the last 4 integrals in (21) which are implicitly neglected by assuming $E[\rho, \boldsymbol{u}, \boldsymbol{b}]$. As for other aspects of the hierarchy, the conservation of mass and entropy are manifest by virtue of the equations of motion for s and ρ having the (divergence) form of local conservation laws. The same applies to momentum, as can be seen from $\partial_t u_i = -\partial_j \left(\sigma_{ij} + \delta_{ij} p\right)$, where the stress tensor $\sigma_{ij} := u_i E_{u_j} + 2b_{ik} E_{b_{jk}}$ and the generalized pressure $p = -e + \rho E_\rho + u_j E_{u_j} + b_{jk} E_{b_{jk}}$ were introduced, e denoting the energy density, and the expansion $\frac{\partial e}{\partial r_j} = E_\rho \partial_j \rho + E_{u_i} \partial_j u_i + E_{b_{ik}} \partial_j b_{ik}$ was used. However, no such conservation law is manifest for the entropy flux $\boldsymbol{a}$.

One of the prospective directions of further investigation is elevating the generalized hydrodynamics given in (22) to an autonomous mesoscopic theory by equipping it (besides the Poisson bracket (21)) with an energy functional, entropy and dissipation potential all given in terms of the state variables (20). Such a theory naturally invites confrontation with experimental observations, numerical simulations (such as those given in [2, 3]) or with other, competing mesoscopic theories for the van der Waals fluid, e.g. the Enskog-Vlasov hydrodynamics investigated in [5].

1.2 van der Waals-Vlasov equation II

In this section we are going to propose another alternative theory of the van der Waals fluid, namely one that bypasses the issues regarding consistency with the equilibrium theory addressed in the previous section at the cost of introducing an additional term proportional to $\frac{\partial f}{\partial v_i}$ to the kinetic equation as well as an entropy that is not a Casimir of the Boltzmann Poisson bracket (10). Let us proceed directly by specifying the kinetic-level entropy as

$$S(f) = -k_B \int d\boldsymbol{r} \int d\boldsymbol{v} f(\boldsymbol{r}, \boldsymbol{v}) \ln(f(\boldsymbol{r}, \boldsymbol{v})) + k_B \int d\boldsymbol{r} n(\boldsymbol{r}) \ln\left(1 - b n(\boldsymbol{r})\right). \tag{23}$$

As already stated, this entropy is not a Casimir of the Boltzmann Poisson bracket. This entails that entropy is not conserved by the kinetics of the theory and irreversibility (with

respect to TRT; encapsulated in the dissipation potential) and dissipation (now present in both the reversible and irreversible terms) play distinct roles. We account for this by taking the thermodynamic potential Φ, rather than energy, as the generating potential of the reversible evolution. On the other hand, since Φ is then obviously conserved by the reversible evolution, any approach of the system towards equilibrium is necessarily driven by the collision term, as is the case both in the Boltzmann equation and in the theory of section 1.1.

Let us keep the remaining ingredients as before and let us derive the corresponding kinetic equation. First, let us express $S_f = k_B \left(\ln(1-bn) - \ln f - \frac{1}{1-bn}\right)$ and

$$\Phi_f = k_B \left(\ln f - \ln(1-bn) + \frac{1}{1-bn}\right) + \frac{\boldsymbol{v}^2}{2mT} + \frac{1}{T}\int d\boldsymbol{r}'\phi(|\boldsymbol{r}-\boldsymbol{r}'|)n(\boldsymbol{r}') - \frac{\mu}{T}. \quad (24)$$

For the reversible part of the equation, this translates into

$$\mathcal{L}\Phi_f = -\frac{v_i}{mT}\frac{\partial f}{\partial r_i} + \left[\frac{1}{T}\frac{\partial \Psi}{\partial r_i} + b\left(\frac{1}{1-bn} + \frac{1}{(1-bn)^2}\right)\frac{\partial n}{\partial r_i}\right]\frac{\partial f}{\partial v_i}.$$

As for the irreversible part, observe that if we now set $f^*(f) := S_f$ and recall the form of the thermodynamic force (6), the terms that only depend on the number density n (and thus have no velocity-dependence) vanish due to the assumption of locality.[2] It follows immediately that the thermodynamic force X has the same form as in the case of the Boltzmann entropy (7) and, using $\sinh(\ln x) = \frac{1}{2}\left(x - \frac{1}{x}\right)$, we arrive at the Boltzmann collision term (8). We conclude that, unlike in the previous section, with our current choice of the constitutive relations (2), there is no modification to the collision term and the effect of the finite size of the particles (expressed via the entropy (23)) is wholly and solely in the additional density-dependent reversible term.

It is quite straightforward to show that the collision term remains invariant if we, instead, set $f^*(f) := -\Phi_f$. Hence, we can write the irreversible term as $\Xi_{f^*}(f, f^*)|_{f^*=S_f} = \Xi_{f^*}(f, f^*)|_{f^*=-\Phi_f}$ and thus declare the thermodynamic potential Φ to be the sole generating potential of this theory. The governing equation can then be written

$$\frac{\partial f}{\partial t} = -\frac{v_i}{m}\frac{\partial f}{\partial r_i} + \left[\frac{\partial \Psi}{\partial r_i} + bT\left(\frac{1}{1-bn} + \frac{1}{(1-bn)^2}\right)\frac{\partial n}{\partial r_i}\right]\frac{\partial f}{\partial v_i} + \mathcal{B}(f), \quad (25)$$

where $\mathcal{B}(f)$ is the Boltzmann collision term (8).

Just as before, let us know proceed with the hydrodynamics. This time, however, we demonstrate the transition for a different set of state variables, namely the moments

$$\begin{aligned} \rho(\boldsymbol{r}) &= \int d\boldsymbol{v} f(\boldsymbol{r},\boldsymbol{v}), & u_i(\boldsymbol{r}) &= \int d\boldsymbol{v} v_i f(\boldsymbol{r},\boldsymbol{v}), \\ b_{ij}(\boldsymbol{r}) &= \int d\boldsymbol{v} v_i v_j f(\boldsymbol{r},\boldsymbol{v}), & c_{ijk}(\boldsymbol{r}) &= \int d\boldsymbol{v} v_i v_j v_k f(\boldsymbol{r},\boldsymbol{v}). \end{aligned} \quad (26)$$

[2]Note that we could have assumed a (more general) non-local thermodynamic force X (resulting in multiple position integrals in (6)) and enforce locality via W in the same manner in which we enforce energy and momentum conservation in the collisions; this would obviously have no (practical) impact on the resulting collision term.

The first three again stand for familiar physical quantities, namely the number density (ρ), the momentum density (u_i) and the stress tensor (b_{ij}). We shall also introduce a special symbol for the contracted third moment $q_i(\boldsymbol{r}) := c_{ijj}(\boldsymbol{r})$, which represents the flow of kinetic energy. As in section 1.1, we assume that all higher-order tensors are zero. Under such assumption, the bracket below can be regarded as being (approximatively) a Poisson bracket.

As mentioned above, the Poisson bracket of the theory is chosen to be the Boltzmann Poisson bracket, except that now we only assume functionals dependent on the fields (26). Hence, substituting the expansion $A_f = A_\rho \frac{\delta\rho}{\delta f} + A_{u_i} \frac{\delta u_i}{\delta f} + A_{b_{ij}} \frac{\delta b_{ij}}{\delta f} + A_{c_{ijk}} \frac{\delta c_{ijk}}{\delta f}$ into (10), we obtain

$$\begin{aligned}
\{A,B\} = \int d\boldsymbol{r}\{&\rho\left(\partial_i(A_\rho)B_{u_i} - \partial_i(B_\rho)A_{u_i}\right) + u_j\left(\partial_i(A_\rho)B_{b_{ij}} - \partial_i(B_\rho)A_{b_{ij}}\right)\\
&+ u_i\left(\partial_j(A_\rho)B_{b_{ij}} - \partial_j(B_\rho)A_{b_{ij}}\right) + b_{ij}\left(\partial_k(A_\rho)B_{c_{ijk}} - \partial_k(B_\rho)A_{c_{ijk}}\right)\\
&+ b_{jk}\left(\partial_i(A_\rho)B_{c_{ijk}} - \partial_i(B_\rho)A_{c_{ijk}}\right) + b_{ik}\left(\partial_j(A_\rho)B_{c_{ijk}} - \partial_j(B_\rho)A_{c_{ijk}}\right)\\
&+ u_i\left(\partial_j(A_{u_i})B_{u_j} - \partial_j(B_{u_i})A_{u_j}\right) + b_{ij}\left(\partial_k(A_{u_i})B_{b_{jk}} - \partial_k(B_{u_i})A_{b_{jk}}\right)\\
&+ b_{ik}\left(\partial_j(A_{u_i})B_{b_{jk}} - \partial_j(B_{u_i})A_{b_{jk}}\right) + c_{ijk}\left(\partial_l(A_{u_i})B_{c_{jkl}} - \partial_l(B_{u_i})A_{c_{jkl}}\right)\\
&+ c_{ilk}\left(\partial_j(A_{u_i})B_{c_{jkl}} - \partial_j(B_{u_i})A_{c_{jkl}}\right) + c_{ilj}\left(\partial_k(A_{u_i})B_{c_{jkl}} - \partial_k(B_{u_i})A_{c_{jkl}}\right)\\
&+ b_{ij}\left(\partial_k(A_{b_{ij}})B_{u_k} - \partial_k(B_{b_{ij}})A_{u_k}\right) + c_{ijk}\left(\partial_l(A_{b_{ij}})B_{b_{kl}} - \partial_l(B_{b_{ij}})A_{b_{kl}}\right)\\
&+ c_{ijl}\left(\partial_k(A_{b_{ij}})B_{b_{kl}} - \partial_k(B_{b_{ij}})A_{b_{kl}}\right) + c_{ijk}\left(\partial_l(A_{c_{ijk}})B_{u_l} - \partial_l(B_{c_{ijk}})A_{u_l}\right)\}.
\end{aligned} \tag{27}$$

The Hamilton's equations for the state variables (26) resulting from this Poisson bracket thus take on the form

$$\begin{aligned}
\frac{\partial\rho}{\partial t} &= -\partial_i\left(\rho\Phi_{u_i} + 2u_j\Phi_{b_{ij}} + 3b_{jk}\Phi_{c_{ijk}}\right)\\
\frac{\partial u_i}{\partial t} &= -\partial_j\left(u_i\Phi_{u_j} + 2b_{ik}\Phi_{b_{jk}} + 3c_{ikl}\Phi_{c_{jkl}}\right)\\
&\quad - \rho\partial_i(\Phi_\rho) - u_j\partial_i(\Phi_{u_j}) - b_{jk}\partial_i(\Phi_{b_{jk}}) - c_{jkl}\partial_i(\Phi_{c_{jkl}})\\
\frac{\partial b_{ij}}{\partial t} &= -\partial_k\left(b_{ij}\Phi_{u_k}\right) - 2\partial_l\left(c_{ijk}\Phi_{b_{lk}}\right)\\
&\quad - u_j\partial_i(\Phi_\rho) - u_i\partial_j(\Phi_\rho) - b_{ik}\partial_j(\Phi_{u_k}) - b_{jk}\partial_i(\Phi_{u_k}) - c_{ikl}\partial_j(\Phi_{b_{kl}}) - c_{jkl}\partial_i(\Phi_{b_{kl}})\\
\frac{\partial c_{ijk}}{\partial t} &= -\partial_l(c_{ijk}\Phi_{u_l})\\
&\quad - b_{jk}\partial_i(\Phi_\rho) - b_{ik}\partial_j(\Phi_\rho) - b_{ij}\partial_k(\Phi_\rho) - c_{ikl}\partial_j(\Phi_{u_l}) - c_{ijl}\partial_k(\Phi_{u_l}) - c_{jkl}\partial_i(\Phi_{u_l}).
\end{aligned} \tag{28}$$

Specifying energy and entropy (and thus the thermodynamic potential Φ) in terms of the state variables (26) would again transform the system (28) into an autonomous mesoscopic theory. All the research possibilities mentioned at the end of the previous section, e.g. comparing the implications of (28) to other Grad-like approaches to phase transitions, such as [5], are at hand. Note, however, an important feature of (28). Unlike classical hydrodynamics as well as the theory presented in section 1.1, our current theory uses neither the field of energy density nor that of entropy density as a state variable. This fact has two important consequences.

The first is that the hydrodynamics presented here preserves the structure of the corresponding kinetic theory and can be seen as its reduced version. This is true both for the hydrodynamic equations (28) and (22). Indeed, the state variables are moments of the state variable (the one-particle distribution function) of the kinetic theory. Their kinematics is directly related to the kinematics of the one particle distribution function[10] and both the energy and the entropy in this hydrodynamic theory are of the same form as their counterparts in the kinetic theory but expressed in terms of the moments (26).

The second consequence is that the van der Waals hydrodynamics presented here is completely free from the local equilibrium assumption. Indeed, the classical hydrodynamics requires that the energy field be in a one-to-one relationship with the entropy field. In other words, the derivative of one of this two fields with respect to the other that is chosen to play the role of one of the state variables is required to be either positive or negative. From the physical point of view, the derivative of the energy field with respect to the entropy field is required to have the physical meaning of the local absolute temperature. The local equilibrium assumption requires that the local entropy depend on the local energy and the local mass in the same way as in the complete thermodynamic equilibrium. The requirement that the derivative of the local entropy with respect to the local energy be the local absolute temperature is thus a weak version of the local equilibrium assumption. Here, we refrain from deploying this assumption in either form.

We are hopeful that further analysis will prove this to be a considerable advantage compared to other hydrodynamic theories, whether aiming at a description of fluids in the hard-sphere approximation or those attempting to describe fluids with a more complex inner structure.

Conclusion

This manuscript aspires to provide a first step on the way towards a microscopic, thermodynamically consistent dynamical theory of the van der Waals fluid based on the mass action law and suitable for the description of phase transitions. Two distinct kinetic equations for the vdW gas are proposed and their respective pros and cons are discussed. Subsequently, the hydrodynamic theories corresponding to each of the equations are derived. These mesoscopic equations are suitable for comparing the predictions of each of the theories with other theoretical, experimental and numerical treatments of phase transitions. Based on the presented arguments, we are hopeful to demonstrate in our future analysis these models as useful dynamical theories of the vdW gas and viable descriptions of phase transition phenomena. Our preliminary theoretical analysis indicates that this is not unlikely, particularly with regard to the approach presented in section 1.2, which leads to a theory free of the local equilibrium assumption. Asymptotic analysis is expected to play a major role in verifying the consistence of the proposed models with the equilibrium theory. Additional theoretical and possibly numerical treatment will facilitate the process of establishing the potential and the limitations of the proposed theories.

References

[1] E. S. Benilov and M. S. Benilov. *Energy conservation and H theorem for the Enskog-Vlasov equation.* Phys. Rev. E **97** (Jun 2018), 062115.

[2] E. S. Benilov and M. S. Benilov. *The enskog–vlasov equation: a kinetic model describing gas, liquid, and solid.* Journal of Statistical Mechanics: Theory and Experiment **2019** (oct 2019), 103205.

[3] T. Chen, C. Zhang, and L.-P. Wang. *Diffuse interface model for a single-component liquid-vapor system.* Phys. Rev. E **107** (Feb 2023), 025104.

[4] D. Enskog. *Kinetische theorie.* Kongl. Vetenskaps Academiens handlingar **63** (1923), 1.

[5] A. Frezzotti and H. Struchtrup. Grad's 13 moments approximation for Enskog-Vlasov equation. volume 2132, 120007, (Aug 2019).

[6] M. Grmela. *Kinetic equation approach to phase transitions.* Journal of Statistical Physics **3** (1971), 347–364.

[7] M. Grmela. *Entropy principle as a restrictive condition on kinetic equations.* Canadian Journal of Physics **59** (1981), 698–707.

[8] M. Grmela and H. C. Öttinger. *Dynamics and thermodynamics of complex fluids. I. Development of a general formalism.* Phys. Rev. E **56** (Dec 1997), 6620–6632.

[9] H. Öttinger. *Beyond Equilibrium Thermodynamics.* Wiley, (2005).

[10] M. Pavelka, V. Klika, O. Esen, and M. Grmela. *A hierarchy of Poisson brackets in non-equilibrium thermodynamics.* Physica D: Nonlinear Phenomena **335** (2016), 54–69.

[11] M. Pavelka, V. Klika, and M. Grmela. *Multiscale Thermo-Dynamics.* De Gruyter, Berlin, Boston, (2018).

[12] H. van Beijeren and M. H. Ernst. *The modified Enskog equation.* Physica **68** (1973), 437–456.

[13] N. G. van Kampen. *Condensation of a classical gas with long-range attraction.* Phys. Rev. **135** (Jul 1964), A362–A369.

[14] A. A. Vlasov. *The vibrational properties of an electron gas.* Soviet Physics Uspekhi **10** (Jun 1968), 721.

[15] L. Waldmann. *Transporterscheinungen in Gasen von mittlerem Druck*, 295–514. Springer Berlin Heidelberg, Berlin, Heidelberg, (1958).

Parameter Estimation in Cyclic Plastic Loading*

Martin Kovanda
kovanma2@fjfi.cvut.cz

study programme: Mathematical Engineering
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Petr Tichavský, Department of Stochastic Informatics
Institute of Information Theory and Automation, CAS

Abstract. Many models have been developed to model the behavior of metallic materials under cyclic plastic loading. However, all of them have to be calibrated using measured data, which typically leads to a non-convex optimization problem with many local minima. In this paper we propose a neural network with a novel loss function that combines estimated parameters and stress responses. Our method outperforms tensor train optimization on synthetic data, but shows similar results on real data.

Keywords: deep learning, neural networks, parameter estimation

Abstrakt. Za účelem modelování chování kovových materiálů při cyklickém plastickém zatížení bylo vyvinuto mnoho modelů, všechny však musí být kalibrovány podle naměřených dat, což obvykle vede k nekonvexnímu optimalizačnímu problému s mnoha lokálními minimy. V této práci představujeme neuronovou síť s novou ztrátovou funkcí, která kombinuje odhadované parametry spolu se stresovými odezvami. Naše metoda překonává optimalizaci tenzorovými vláčky na syntetických datech, na reálných datech však vykazuje podobné výsledky.

Klíčová slova: hluboké učení, neuronové sítě, odhad parametrů

1 Introduction

There has always been a need to model the behavior of solid materials under load. A well-performing model would speed up the design process of individual machine components. Throughout history, there have been many ways to study material behavior, such as static or impact loading tests. However, numerous bridge and railroad failures increased the need to study material behavior under repeated loading. To this end, cyclic plastic loading became one of the top priorities in materials research.

The behavior of materials under cyclic loading is a complex problem that has challenged researchers for decades. One phenomenon observed during cyclic loading is the Bauschinger effect. When a metallic specimen is stretched beyond its elastic range, the stress begins to cause permanent microscopic changes. The material begins to adapt to the new stress state while reducing its ability to withstand compressive stress.

*This work was supported by MEYS CR under grant No. LTA USA 18199 and by the Czech Science Foundation through the project No. 22-11101S.

Each material model must first be calibrated using data from a real cyclic loading experiment. Each experiment is controlled by the total deformation $\epsilon(t)$, which is a function of time. The measurable output is the stress $S(\epsilon)$ measured in MPa, which is a function of the history of the total deformation and is considered independent of the speed of the experiment. The total deformation is a sum of the elastic and plastic deformation $\epsilon_{\mathrm{e}}(t)$ and $\epsilon_{\mathrm{p}}(t)$ respectively. Since the elastic deformation can be easily subtracted, it is more convenient for the next purposes to work with only the plastic deformation $\epsilon_{\mathrm{p}}(t)$. For this reason, we will consider the stress as a function of only the plastic deformation history, $S(\epsilon_{\mathrm{p}})$. The elastic properties of the material, defined by the Young's modulus E, are determined analytically.

Many hardening models have been developed to address various identified phenomena, such as the Bauschinger effect. A prominent single yield surface model with nonlinear hardening is the Armstrong-Frederick model (1966) [1]. Building upon similar principles, the MAFTr model introduced by Dafalias and Feigenbaum in 2011 [3] adds linear hardening to one of the backstress components and further refines the behavior under multi-axial loading.

In this paper, the analytical solution of the uni-axial MAFTr model presented by Marek et al. (2022) [6] is used. However, the linear hardening has been removed for simplicity. This analytical model $M_{\boldsymbol{\theta}}$ represents a way to predict the stress response of a metallic specimen given the plastic deformation history $\epsilon_{\mathrm{p}}(t)$ and its material vector parameter $\boldsymbol{\theta} \in \mathbb{R}_+^N$. The composition of $\boldsymbol{\theta}$ is shown in Table 1. The vectors $\mathbf{c}$ and $\mathbf{a}$ have the same arbitrary dimension. With a higher dimension, the model may be more accurate, but at the cost of increased complexity. In this paper we only use a dimension of 4, which gives $\boldsymbol{\theta} \in \mathbb{R}_+^{11}$.

Parameter	**Unit**	**Description**
k_0	MPa	Initial yield strength
κ_1	MPa	Adjustment of the rate of isotropic hardening
κ_2	MPa^{-1}	Inverted asymptotic limit of isotropic hardening
c_i	-	Adjustment of the evolution rates of the backstress components
a_i	MPa	Asymptotic limits of the backstress components

Table 1: Parameters of analytical model developed by Marek et al. and their descriptions.

The analytical model can never predict the material behavior exactly. The task is therefore to find an optimal $\boldsymbol{\theta}^*$ that describes it as accurately as possible, i.e. to find

$$\boldsymbol{\theta}^* := \underset{\boldsymbol{\theta} \in \mathbb{R}_+^{11}}{\arg\min}\, L_2\big(\big|\mathbf{S}_m - M_{\boldsymbol{\theta}}(\epsilon_{\mathrm{p}})\big|\big) \tag{1}$$

for a known ϵ_{p} corresponding to the measured experimental setup, where $\mathbf{S}_m$ refers to the measured stress response. Estimating $\boldsymbol{\theta}^*$ is a non-trivial problem since $L_2(|\mathbf{S}_m - M_{\boldsymbol{\theta}}(\epsilon_{\mathrm{p}})|)$ is not a convex function. Current approaches often use a random search to determine the initial point for the Nelder–Mead method [7] (hereafter referred to as "simplex"). This method is a non-gradient optimization technique that iteratively refines potential solutions until an optimal result is achieved. However, this approach is time-consuming and usually finds a suboptimal solution that is far from the global minimum.

In this paper we develop several types of neural network estimators $\widehat{\boldsymbol{\theta}}_{\mathrm{NN}}$. A sufficiently close estimate can then be used as an initial point for another optimization method, such as simplex, which would find a $\boldsymbol{\theta}$ close to a local minimum. This would significantly reduce the time complexity compared to a random search that yields similar stress estimates.

The best performing neural networks are then compared to the recently published tensor train optimization method, TTOpt, introduced by Sozykin et al.[8], which is a general non-gradient method for approximating multivariable functions.

2 Data Preparation

In the real experiment, the stress response was recorded at a sampling rate of 10Hz for 4 hours. Since the stress does not depend on the speed of the experiment, it may be beneficial to first downsample this data to reduce computational complexity without significantly reducing the information contained. This process mimics increasing the speed of the experiment, however, it needs to preserve points of reversals as they in fact define the experiment. To make each segment equally informative, in this paper we downsample each segment of plastic loading differently, so that each segment consists of exactly $N = 14$ increments resulting in 15 samples including the edge points.

Let $\epsilon_0 = 0$ stand for the plastic deformation in the beginning of the experiment and let $\epsilon_1^{(\mathrm{r})}, \ldots, \epsilon_K^{(\mathrm{r})}$ represent the plastic deformation in the end of each segment, where $K = 43$ is the number of all segments. Then the i-th segment is interpolated in plastic deformations $\epsilon_i^{(1)}, \ldots, \epsilon_i^{(N-1)}$, where $N = 14$ represents the total number of increments in each segment. Most of the useful information is expected to be at the beginning of each load segment, where the stress level changes more rapidly. Therefore, these points are defined as

$$\epsilon_i^{(j)} := \epsilon_{i-1}^{(\mathrm{r})} + \sum_{k=1}^{j} \delta_k, \quad \forall i \in \{1, \ldots, K\},\ \forall j \in \{1, \ldots, N-1\}, \tag{2}$$

using geometrical sequence of increments $\delta_1, \ldots, \delta_N$ generated so that

$$\epsilon_i^{(\mathrm{r})} = \epsilon_{i-1}^{(\mathrm{r})} + \sum_{i=1}^{N} \delta_i \qquad \text{and} \qquad \delta_{k+1} = \sqrt[N-1]{R}\,\delta_k,\ \forall k \in \{1, \ldots, N-1\} \tag{3}$$

for a chosen parameter $R = 20$. This setup is designed to make the ratio between the first and the last increment in each cycle equal to R, thus $\frac{\delta_N}{\delta_1} = R$. Figure 1 shows the resulting deformations from the measured experiment $\boldsymbol{\epsilon}_{\mathrm{p}}^{(\mathrm{exp})}$, defined as

$$\boldsymbol{\epsilon}_{\mathrm{p}}^{(\mathrm{exp})} := \big(0, \epsilon_1^{(1)}, \ldots, \epsilon_1^{(N-1)}, \epsilon_1^{(\mathrm{r})}, \epsilon_2^{(1)}, \ldots, \epsilon_2^{(N-1)}, \epsilon_2^{(\mathrm{r})}, \ldots, \epsilon_K^{(\mathrm{r})}\big). \tag{4}$$

After preparing the sequence of plastic deformations and their associated stress responses, it is possible to estimate the parameter $\boldsymbol{\theta}$. For both neural networks and TTOpt, it is necessary to first select an a priori distribution for $\boldsymbol{\theta}$. In the case of neural networks, this allows the creation of a training set of stress responses and corresponding parameters (see Section 3). TTOpt, on the other hand, needs a region defined by the Cartesian product of intervals $[\theta_{i,\min}, \theta_{i,\max}]$ for each parameter θ_i in which the optimal $\boldsymbol{\theta}$ is to be found, together with a properly chosen sampling for each of these intervals.

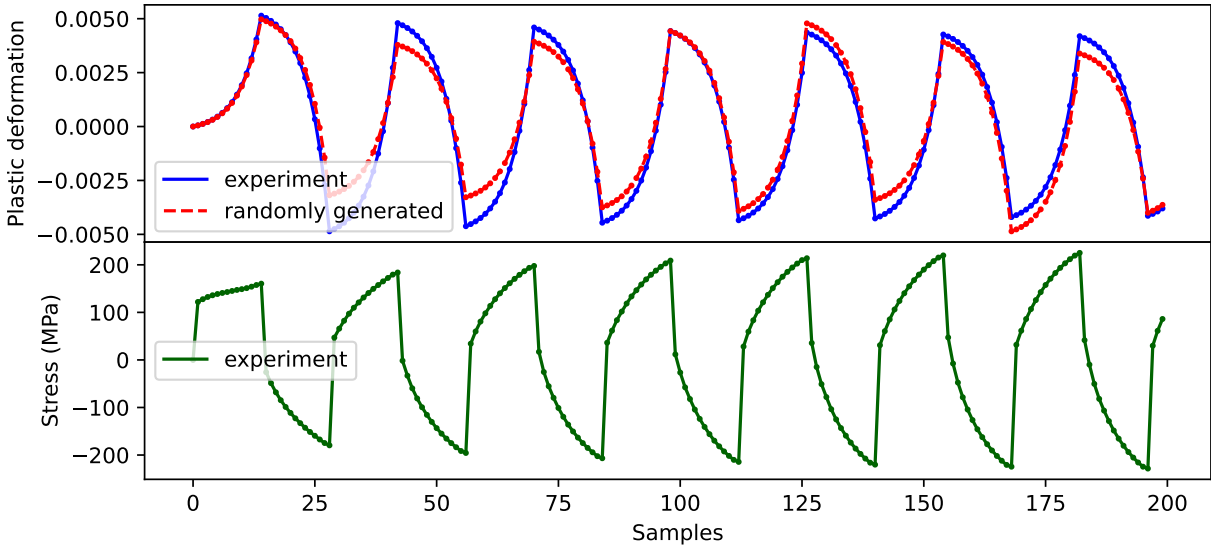


Figure 1: Top: Example of the first 200 interpolated points of deformation taken to cover each load segment by 15 points using a geometric sequence of increments defined in Equation 3. Bottom: Interpolated stress from the measured data corresponding to the interpolated points of deformation.

To cover the entire handpicked interval, a uniform distribution is chosen for each parameter θ_i separately, see Table 2. All these ranges have been chosen to cover the most commonly used materials and can be easily adjusted if needed. There are 2 additional conditions. The parameters a_i are generated so that their sum would be in range of $[150, 350]$. Second, since the pairs (c_i, a_i), $i \in \{1, \dots, 4\}$ are commutative, they are generated with the condition $c_1 \geq c_2 \geq c_3 \geq c_4$ to make the training objective unique. Unsorted parameters would make it harder for neural networks to predict correct values because their order would be inconsistent. For example, an optimal solution with permuted (c_i, a_i) pairs could be considered incorrect.

	k_0	κ_1	κ_2^{-1}	$\log(c_1)$	$\log(c_{2,3,4})$	$a_{1,2,3,4}$
min	15	100	30	$\log(1000)$	$\log(50)$	0
max	250	10000	150	$\log(10000)$	$\log(2000)$	350

Table 2: Range of the a priori uniform distribution for each (transformed) parameter, given the conditions $\sum a_i \in [150, 350]$ and $c_1 \geq c_2 \geq c_3 \geq c_4$. The symbol log stands for the natural logarithm.

In practice, the first condition is realized by first generating $\widetilde{a} \sim U[150, 350]$ and $a'_1, \dots, a'_4 \sim U[0, 1]$. Then the desired parameters $a_1, \dots, a_4$ are calculated as

$$a_i := \frac{a'_i}{\sum_{j=1}^{4} a'_j} \widetilde{a}, \qquad \forall i \in \{1, \dots, 4\}. \tag{5}$$

This way mimics generating $a_i \sim U[0, 350]$ while satisfying the condition $\sum a_i \in [150, 350]$. The second condition is solved by simply sorting the c_i parameters.

3 Dataset

Training neural networks requires large amounts of data. Running real experiments on such a scale is not feasible. The analytical model $M_{\boldsymbol{\theta}}$ provides a fast way how to obtain approximate stress responses in a cyclic plastic loading experiment given the parameters $\boldsymbol{\theta}$ and plastic deformation ϵ_P. For this reason, the dataset is created by generating $\boldsymbol{\theta}$ from a designed random distribution and then obtaining stress responses using the analytical model $M_{\boldsymbol{\theta}}$.

In this paper 2 datasets are created. The first dataset is created to train neural networks specifically for the plastic deformation measured in the real experiment $\boldsymbol{\epsilon}_{\mathrm{p}}^{(\mathrm{exp})}$. The second dataset is then generalized so that the neural networks would be able to perform parameter estimation for any real experiment, i.e., an experiment with any plastic deformation setup.

Let $\mathbf{P}_{11}$ represent the a priori distribution for $\boldsymbol{\theta}$ described in Section 2. The first dataset $\mathbf{D}_1$ consists of pairs $(S_i, \boldsymbol{\theta}_i)$ and is created as

$$\mathbf{D}_1 := \left\{ \left(\mathbf{M}_{\boldsymbol{\theta}_i}\left(\boldsymbol{\epsilon}_{\mathrm{p}}^{(\mathrm{exp})}\right),\ \boldsymbol{\theta}_i \right),\ i \in \{1, \ldots, L\} \right\}, \quad \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L \overset{\mathrm{iid}}{\sim} \mathbf{P}_{11}, \tag{6}$$

where $L = 10^6$ is the chosen length of the dataset. The plastic deformation does not need to be a part of the dataset, as it remains constant across all generated data. In neural networks, this is not an issue because their biases can effectively deal with such constants.

The second dataset is generated to develop models capable of parameter estimation independent of the experimental deformation setting. In this case the newly generated dataset $\mathbf{D}_2$ must also include the plastic deformation, since it is different in all the generated data. These deformations need to be generated to mimic real experiments. Since the stress depends only on the deformation and the speed of the experiment is arbitrary, it is sufficient to first generate the plastic deformation only in the end of each load segment $\epsilon_1^{(\mathrm{r})}, \ldots, \epsilon_K^{(\mathrm{r})}$, where $K = 43$ stands for the number of segments. The odd (elongating) deformations are generated using a handpicked a priori uniform distribution $U[0.003, 0.005]$, while the even (compressing) deformations are taken from $U[-0.005, -0.003]$. The data points within each segment are then generated using Equation 2. An example of a randomly generated set of data points is shown in Figure 1.

Let $\mathbf{U}_\epsilon$ represent the combined distribution of all data samples generated using the procedure described above, i.e.

$$\boldsymbol{\epsilon}_{\mathrm{p}} := \left(0, \epsilon_1^{(1)}, \ldots, \epsilon_1^{(N-2)}, \epsilon_1^{(\mathrm{r})}, \epsilon_2^{(1)}, \ldots, \epsilon_2^{(N-2)}, \epsilon_2^{(\mathrm{r})}, \ldots, \epsilon_K^{(\mathrm{r})}\right), \qquad \forall \boldsymbol{\epsilon}_{\mathrm{p}} \sim \mathbf{U}_\epsilon, \tag{7}$$

where $N = 15$ is the number of deformations in each segment and $K = 43$ stands for the number of segments. Then the second dataset $\mathbf{D}_2$ consists of triplets $(S_i, \boldsymbol{\epsilon}_{P,i}, \boldsymbol{\theta}_i)$ and is created as

$$\mathbf{D}_2 := \left\{ \left(\mathbf{M}_{\boldsymbol{\theta}_i}\left(\boldsymbol{\epsilon}_{P,i}\right),\ \boldsymbol{\epsilon}_{P,i},\ \boldsymbol{\theta}_i \right),\ i \in \{1, \ldots, L\} \right\}, \quad \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L \overset{\mathrm{iid}}{\sim} \mathbf{P}_{11}, \quad \boldsymbol{\epsilon}_{P,1}, \ldots, \boldsymbol{\epsilon}_{P,L} \overset{\mathrm{iid}}{\sim} \mathbf{U}_\epsilon, \tag{8}$$

where $L = 10^6$ is the chosen length of the dataset.

Similar to the training datasets $\mathbf{D}_1$ and $\mathbf{D}_2$, corresponding validation datasets $\mathbf{D}_1^V$, $\mathbf{D}_2^V$ composed of 20000 pairs and test datasets $\mathbf{D}_1^T$, $\mathbf{D}_2^T$ with 1024 pairs are created for evaluation purposes.

4 Neural Network Architectures

In recent years, neural networks have started to dominate over previous methods due to their ability to learn non-trivial features from the training dataset using a gradient-based self-propagation method. However, there are still challenges in proposing an appropriate architecture and finding optimal hyperparameters. This often involves training multiple architectures on the training dataset and selecting the best performing one on the validation dataset.

Each architecture takes as an input a matrix of shape 1×603 for the data set $\mathbf{D}_1$ and 2×603 for $\mathbf{D}_2$ because the second data set also contains plastic deformation. To bring the variance closer to 1, the first layer of each architecture is the batch normalization, which converts the input data separately for each channel (stress and plastic deformation).

Each parameter θ_i has a different order of magnitude. For this reason, it would be difficult to train neural networks to predict them directly, since the L_2 loss function used would generally favor some parameters over others. To avoid this potential problem, each $\boldsymbol{\theta}$ is then normalized element-wise based on means and variances determined by the distribution $\mathbf{P}_{11}$. In practice, these means and variances are estimated based on 10^6 realizations of $\mathbf{P}_{11}$.

Let $\boldsymbol{\theta}^{(\mathcal{N})}$ represent the parameter $\boldsymbol{\theta}$ normalized element-wise using the mean and variance estimates. The output of each model is a vector of length 11 representing the parameter $\boldsymbol{\theta}^{(\mathcal{N})}$. After scaling, there is no easy way to prevent the networks from predicting negative values for individual parameters. For this reason, post-processing must include replacing any negative estimated parameter with a small number $\eta = 10^{-9}$ chosen for numerical stability.

The performance of the developed architectures is measured by 2 metrics. The first one is the $L_{\boldsymbol{\theta}}$, defined as

$$L_{\boldsymbol{\theta}} := \frac{1}{11} \sum_{i=1}^{11} \left(\theta_i^{(\mathcal{N})} - \widehat{\theta}_i^{(\mathcal{N})}\right)^2. \tag{9}$$

The normalized parameter $\boldsymbol{\theta}^{(\mathcal{N})}$ is used to deal with different scales between individual parameters. The second and more important metric is the $L_{\mathbf{S}}$, which measures the average difference between the reference stress $\mathbf{S}$ and the predicted stress, i.e.

$$L_{\mathbf{S}} := \frac{1}{603} \sum_{i=1}^{603} \left(S_i - \widehat{S}_i\right)^2, \qquad \widehat{\mathbf{S}} := M_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{\epsilon}_{\mathrm{p}}). \tag{10}$$

This metric indicates how far the predicted stress is from the reference stress, taking into account the existence of many local minima in $L_2(|\mathbf{S}_m - M_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}_{\mathrm{p}})|)$.

In the following experiments, the shape (number of layers, number of neurons in each layer, etc.) is first found using a manually directed random search. At this stage, all networks are trained using the AdamW optimizer created by Loshchilov and Hutter (2017) [5], with a learning rate of 10^{-3} and a weight decay of 10^{-2}. The batch size is set to 64. Each network is then evaluated on a validation dataset consisting of 20000 data generated similarly to $\mathbf{D}_1$ and $\mathbf{D}_2$. Based on this evaluation, one network is selected for the second phase. Here, different optimizer settings and batch sizes are tried to further improve the performance of the selected network.

4.1 Feed-Forward Networks (FFN)

One of the simplest architectures is a neural network based solely on feed-forward layers. Each hidden layer is followed by a ReLU activation function. Since $\boldsymbol{\theta}^{(\mathcal{N})}$ can be negative, no activation function is used after the last layer. ReLU is used because it is the simplest nonlinear activation function commonly used for neural networks.

The set of hyperparameters consists of the number of layers and the number of neurons in each layer. First, the layer structure is found by random search using a set of $2^5, \ldots, 2^{10}$ with decreasing number of neurons in each successive layer. The performance of 3 selected networks is shown in Table 3. Numerical experiments indicate that having less than 3 layers leads to underperforming networks. Meanwhile, adding more layers does not seem to significantly improve performance. For this reason, network number 2 is chosen for the following experiments.

id	fully-connected layers	$L_{\boldsymbol{\theta}}$	$L_{\mathbf{S}}$
1	[512, 256, 128, 64]	0.321	**61.100**
2	[512, 256, 128]	**0.318**	63.092
3	[256, 128]	0.321	112.010

Table 3: Metrics of 3 chosen feed-forward networks on validation dataset $\mathbf{D}_1^V$.

The next step is to find a good optimizer setting. Experiments show that using AdamW instead of Adam optimizer produces better performing networks. The best found setting was using a batch size of 128 and beta parameters $(0.9, 0.99)$, even though the value of $L_{\mathbf{S}}$ is similar to that in Table 3. This may indicate that the initial optimizer setting was already strong.

4.2 Recurrent Neural Networks (RNN)

Recurrent neural networks are widely used for time series analysis. The nature of the datasets $\mathbf{D}_1$ and $\mathbf{D}_2$ suggests that RNNs might have a better performance than CNNs. Since the simple recurrent unit tends to perform worse for longer inputs, in this paper we instead use both Gated Recurrent Unit (GRU) [2] and Long Short-Term Memory (LSTM) [4] units. These commonly used architectures are designed to overcome the problem of processing long sequences and could therefore potentially perform better on cyclic plastic loading data.

LSTM and GRU both have as hyperparameters number of layers and hidden size, which indicates the dimensionality of the hidden state and directly affects the capacity and complexity of the model. As shown in Table 4, both architectures appear to have a similar performance.

Similar to FFN, random search does not significantly improve network performance for different optimizer setups. The only improvement comes from increasing the batch size to 128 instead of 64. Both GRU and LSTM winning networks outperform the FFN baseline. However, since GRU performs better than LSTM while being significantly faster and simpler, only GRU is considered in the following experiments.

id	unit type	GRU/LSTM layers	hidden size	$L_{\boldsymbol{\theta}}$	$L_{\mathbf{S}}$
1	GRU	6	128	**0.169**	**31.261**
2	GRU	8	128	0.167	32.217
3	LSTM	6	64	0.189	33.977
5	GRU	4	128	0.173	37.762
7	LSTM	4	128	0.195	41.569
8	LSTM	6	128	0.185	43.329

Table 4: Metrics of 3 chosen LSTM and GRU networks on validation dataset $\mathbf{D}_1^V$.

4.3 Dataset $\mathbf{D}_2$

The second data set consists of both stress response and plastic deformation, as it is designed to train networks capable of estimating parameters for various deformation setups. This makes the second data set more challenging. By analogy to a random search performed on the $\mathbf{D}_1$ dataset, similar architectures also perform well on the more complex $\mathbf{D}_2$ dataset. Table 5 shows that GRU outperforms the baseline and gives better results in terms of $L_{\boldsymbol{\theta}}$ and $L_{\mathbf{S}}$. The winning GRU network has 8 GRU layers, while the best one for dataset $\mathbf{D}_1$ had 6 layers.

id	architecture	GRU layers	hidden size	fully-connected layers	$L_{\boldsymbol{\theta}}$	$L_{\mathbf{S}}$
1	GRU	8	128	-	0.144	**29.224**
2	GRU	6	128	-	0.164	29.344
3	FFN	-	-	(256, 256, 128)	0.324	66.388
4	FFN	-	-	(512, 256, 128, 64, 32)	0.321	70.487

Table 5: Metrics of 2 chosen GRU and FFN networks on validation dataset $\mathbf{D}_2^V$.

4.4 Enhancement of Loss Function

Using only the L_2 loss function between the predicted $\widehat{\boldsymbol{\theta}}_{\mathcal{N}}$ and the reference $\boldsymbol{\theta}_{\mathcal{N}}$ proved to be sufficient to make relatively close estimates. However, since the final objective is to minimize the distance between the reference stress $\mathbf{S}$ and the predicted stress $\widehat{\mathbf{S}} := M_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{\epsilon}_{\mathrm{p}})$, the training can be further improved by using a linear combination between the parameter loss and the stress loss, i.e.

$$L(\widehat{\boldsymbol{\theta}}_{\mathcal{N}}, \boldsymbol{\theta}_{\mathcal{N}}, \widehat{\mathbf{S}}, \mathbf{S}) := kL_2\big(\widehat{\boldsymbol{\theta}}_{\mathcal{N}}, \boldsymbol{\theta}_{\mathcal{N}}\big) + \alpha(1-k)L_2\big(\widehat{\mathbf{S}}, \mathbf{S}\big), \qquad k \in [0,1],\ \alpha \in \mathbb{R}_+. \tag{11}$$

Experiments show that training the networks using this loss improves the performance of the models as they are now trained to make closer stress estimates while also trying to predict similar $\boldsymbol{\theta}$, see Figure 2. The parameter $\alpha = 30$ is chosen to minimize the difference between $L_2\big(\widehat{\boldsymbol{\theta}}_{\mathcal{N}}, \boldsymbol{\theta}_{\mathcal{N}}\big)$ and $L_2\big(\widehat{\mathbf{S}}, \mathbf{S}\big)$ at the beginning of training.

This approach requires being able to compute backpropagation over the analytic model $M_{\boldsymbol{\theta}}$, which significantly increases computational complexity for calculating $L_2\big(\widehat{\mathbf{S}}, \mathbf{S}\big)$. The optimal k seems to be $k = 3$ for FFN and $k = 6$ for GRU to balance $L_{\boldsymbol{\theta}}$ and $L_{\mathbf{S}}$, see Figure 2.

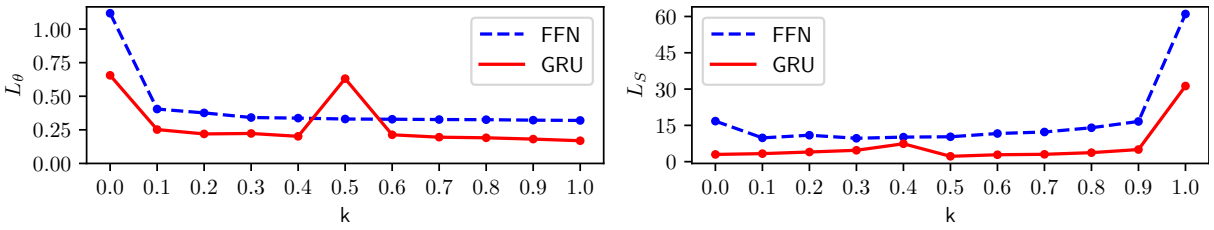


Figure 2: Comparison of $L_{\boldsymbol{\theta}}$ and $L_{\mathbf{S}}$ metrics for FFN and GRU trained with $L(\widehat{\boldsymbol{\theta}}_{\mathcal{N}}, \boldsymbol{\theta}_{\mathcal{N}}, \widehat{\mathbf{S}}, \mathbf{S})$ for different k values.

4.5 Comparison to TTOpt

After the final GRU architectures are selected using the validation datasets $\mathbf{D}_1^V$ and $\mathbf{D}_2^V$, they are then evaluated on the test datasets $\mathbf{D}_1^T$ and $\mathbf{D}_2^T$. Since these networks are only meant to provide a starting point for other optimization methods, the final Table 6 also contains metrics of the predictions optimized by simplex. Simplex optimization is only used for stress prediction, since the optimal $\boldsymbol{\theta}^*$ for the measured experiment would not be known in advance.

architecture	dataset	$L_{\boldsymbol{\theta}}$	$L_{\mathbf{S}}$	$L_{\boldsymbol{\theta}}$ - after simplex	$L_{\mathbf{S}}$ - after simplex
GRU	$\mathbf{D}_1^T$	**0.204**	**3.423**	**0.242**	**0.004**
FFN	$\mathbf{D}_1^T$	0.331	7.211	0.408	0.024
TTOpt	$\mathbf{D}_1^T$	4.942	241.781	4.370	4.593
GRU	$\mathbf{D}_2^T$	**0.186**	**3.479**	**0.216**	**0.004**
FFN	$\mathbf{D}_2^T$	0.373	7.248	0.437	0.037
TTOpt	$\mathbf{D}_2^T$	4.758	266.940	4.077	3.806

Table 6: Metrics of selected GRU and FFN networks compared to TTOpt on test datasets $\mathbf{D}_1^T$ and $\mathbf{D}_2^T$.

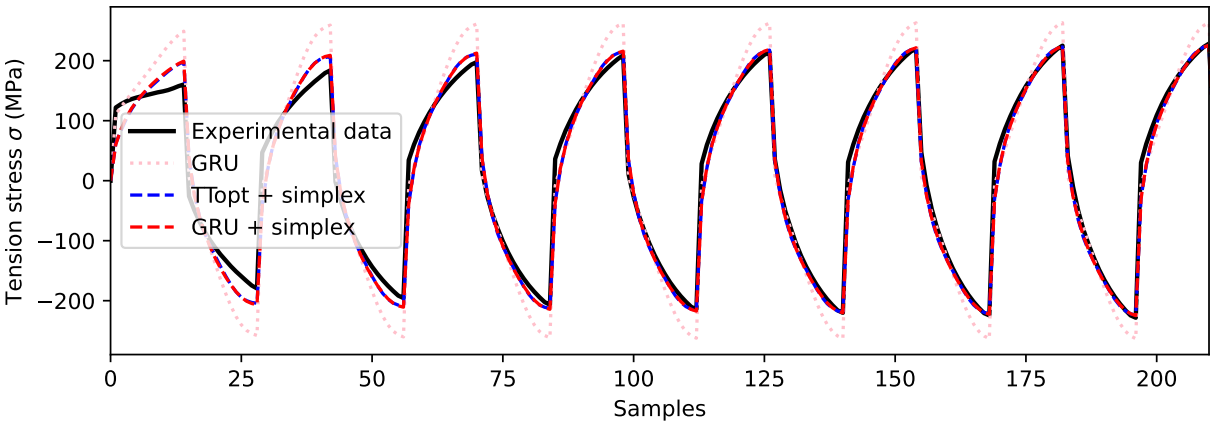


Figure 3: Comparison between the real experiment stress and stress calculated using the estimated parameters $\boldsymbol{\theta}$.

Compared to TTOpt, GRU seems to give a better prediction on synthetic data. How-

ever, as shown in Figure 3, these two approaches give very similar estimates on real experiment. This may indicate that having only synthetic data without additional augmentation may not be sufficient to train neural networks capable of optimal estimation on real data. Future research would be needed to determine what kind of augmentation could possibly help to make better estimates.

5 Conclusion

Parameter estimation for a cyclic plastic loading model is challenging because it leads to the minimization of a non-convex function with many local minima. An effective strategy to solve this problem is the deployment of neural networks, specifically trained on a synthetic dataset of stress-parameter pairs. Among these networks, the recurrent architectures, in particular GRU and LSTM, have shown great promise.

Neural network training can be further improved by using a loss function based on combining a loss of both the parameter estimation and the subsequent stress response prediction. GRU trained in this way showed a high performance on the synthetic dataset compared to the tensor train optimization method TTOpt. However, on real data, the performance of GRU and TTOpt seems to be similar.

References

[1] P. J. Armstrong, C. Frederick, et al. *A mathematical representation of the multiaxial Bauschinger effect*, volume 731. Berkeley Nuclear Laboratories Berkeley, CA, (1966).

[2] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, (2014).

[3] Y. Dafalias and H. Feigenbaum. *Biaxial ratchetting with novel variations of kinematic hardening.* International Journal of Plasticity - INT J PLASTICITY **27** (04 2011), 479–491.

[4] S. Hochreiter and J. Schmidhuber. *Long short-term memory.* Neural computation **9** (12 1997), 1735–80.

[5] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, (2019).

[6] R. Marek, S. Parma, V. Klepač, and H. P. Feigenbaum. *A quick calibration tool for cyclic plasticity using analytical solution.* Engineering Mechanics **27/28** (May 2022), 249 – 252.

[7] J. A. Nelder and R. Mead. *A simplex method for function minimization.* Computer Journal **7** (1965), 308–313.

[8] K. Sozykin, A. Chertkov, R. Schutski, A.-H. Phan, A. Cichocki, and I. Oseledets. TTOpt: A Maximum Volume Quantized Tensor Train-based Optimization and its Application to Reinforcement Learning, (September 2022). arXiv:2205.00293 [cs, math].

A Lattice Boltzmann Approach to Mathematical Modeling of Myocardial Perfusion*

Jan Kovář
kovarj29@fjfi.cvut.cz

study programme: Mathematical Engineering
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. A mathematical model of myocardial perfusion based on the lattice Boltzmann method (LBM) is proposed and its applicability is investigated in both healthy and diseased cases. The myocardium is conceptualized as a porous material in which the transport and mass transfer of a contrast agent in blood flow is studied. The results of myocardial perfusion obtained using LBM in 1D and 2D are confronted with previously reported results in the literature and the results obtained using the mixed-hybrid finite element method. Since LBM is not suitable for simulating flow in heterogeneous porous media, a simplified and computationally efficient 1D-analog approach to 2D diseased case is proposed and its applicability discussed.

A biophysical model of myocardial perfusion coupled with a model of the transport of CA in the vessels and myocardial tissue, see e.g. book chapter [26] and references therein, combined with the data acquired during perfusion magnetic resonance imaging (MRI) represents a great potential in providing more information about the proportion of microvascular disease vs. pathology in large coronaries. Simulating the tissue perfusion is a computationally intensive task [28]. The main goal of the present work is to advance the biophysical myocardial perfusion modeling toward the applicability in the real clinical setup. Specifically, the paper proposes a computationally efficient approach that would possibly allow to connect the model with data directly during the perfusion MRI exam. Two types of acceleration are investigated: a) applicability of the lattice Boltzmann method (LBM) for solving the mathematical model of myocardial perfusion in a 2D homogeneous porous medium, and b) spatial reduction of the LBM model to 1D and its suitability to simulate the perfusion both in a homogeneous and heterogeneous (by coupling two LBM-based 1D problems — "1D-analog") porous media. The former aims to represent the healthy heart or the myocardium with a homogeneously decreased perfusion (such as, e.g., in microvascular disease affecting the whole heart), the latter simulates a perfusion defect (such as in epicardial coronary artery disease).

The results obtained by the proposed LBM model for the healthy cases in 1D and 2D are compared to those reported by Cookson et al. [9]. Secondly, the proposed LBM-based 1D-analog to the 2D problem, simulating the perfusion in a heterogeneous porous medium, is assessed. The resulting simulated temporal profiles of CA concentrations in the healthy tissue and perfusion defect are then compared to the results of the 2D perfusion problem obtained by the mixed-hybrid finite element method (MHFEM) [18] which serves as the reference numerical method in this study.

*This work has been supported by the grant No. NV19-08-00071

Keywords: Lattice Boltzmann method, mixed-hybrid finite element method, myocardial perfusion, magnetic resonance imaging, advection-diffusion problem, contrast agent transport

Abstrakt. V rámci tohoto článku je navržen matematický model perfuze myokardu založený na mřížkové Boltzmannově metodě (LBM) a je zkoumána jeho použitelnost u zdravých i nemocných pacientů. Myokard je koncipován jako porézní prostředí, ve kterém je studován transport kontrastní látky v krevním proudění, včetně přestupu hmoty přes cévní stěnu. Výsledky perfuze myokardu získané pomocí LBM v 1D a 2D jsou konfrontovány s dříve uvedenými výsledky v literatuře a s výsledky získanými pomocí smíšené hybridní metody konečných prvků. Protože LBM není vhodná metoda pro simulaci proudění v heterogenních porézních prostředích, je navržen zjednodušený a výpočetně efektivní 1D model k 2D modelu nemocných a diskutována jeho použitelnost.

Biofyzikální model perfuze myokardu spojený s modelem transportu CA v cévách a tkáni myokardu, viz např. kapitola [26] a odkazy v ní, v kombinaci s daty získanými během perfuzního MRI představuje velký potenciál pro získání více informací o poměru mikrovaskulárního onemocnění a patologie ve velkých koronárních cévách. Simulace perfuze tkání je výpočetně náročná úloha [28]. Hlavním cílem této práce je posunout biofyzikální modelování perfuze myokardu směrem k použitelnosti v reálném klinickém prostředí. Konkrétně je v práci navržen výpočetně efektivní přístup, který by případně umožnil propojit model s daty přímo během perfuzního MRI vyšetření. Zkoumají se dva typy zrychlení: a) použitelnost mřížkové Boltzmannovy metody (LBM). pro řešení matematického modelu perfuze myokardu ve 2D homogenním porézním prostředí a b) prostorová redukce modelu LBM na 1D a jeho vhodnost pro simulaci perfuze v homogenním i heterogenním (spojením dvou 1D úloh založených na LBM — „1D analog") porézním prostředí. První z nich má za cíl reprezentovat zdravé srdce nebo myokard s homogenně sníženou perfuzí (jako např. u mikrovaskulárního onemocnění postihujícího celé srdce), druhá simuluje perfuzní defekt (jako např. u epikardiální ischemické choroby srdeční).

Výsledky získané pomocí navrženého modelu LBM pro případy zdravé tkáně v 1D a 2D jsou porovnány s výsledky, které uvádí Cookson et al. [9]. Dále je navrhovaný 1D analog srovnán s výsledky 2D úlohy simulující perfuzi v heterogenním porézním prostředí. Výsledné simulované časové profily koncentrací kontrastní látky ve zdravé tkáni a v defektu perfuze jsou pak porovnány s výsledky 2D úlohy perfuze získanými smíšenou hybridní metodou konečných prvků [18], která v této studii slouží jako referenční numerická metoda.

Klíčová slova: mřížková Boltzmannova metoda, smíšená hybridní metoda konečných prvků, perfuze myokardu, vyšetření magnetickou rezonancí, advekčně-difuzní úloha, transport kontrastní látky

Full paper: R. Fučík, J. Kovář, K. Škardová, O. Polívka, R. Chabiniok: A Lattice Boltzmann Approach to Mathematical Modeling of Myocardial Perfusion. Under review in International Journal for Numerical Methods in Biomedical Engineering.

References

[1] H. M. Arthur, P. Campbell, P. J. Harvey, M. McGillion, P. Oh, E. Woodburn, and C. Hodgson. *Women, cardiac syndrome X, and microvascular heart disease.* Canadian journal of cardiology **28** (2012), S42–S49.

[2] L. Axel. *Tissue mean transit time from dynamic computed tomography by a simple deconvolution technique.* Investigative radiology **18** (1983), 94–99.

[3] R. Chabiniok, B. Burtschell, D. Chapelle, and P. Moireau. *Dimensional reduction of a poromechanical cardiac model for myocardial perfusion studies.* Applications in Engineering Science (2022), 100121.

[4] D. Chapelle, J.-F. Gerbeau, J. Sainte-Marie, and I. Vignon-Clementel. *A poroelastic model valid in large strains with applications to perfusion in cardiac modeling.* Computational Mechanics **46** (2010), 91–101.

[5] D. Chapelle, J.-F. Gerbeau, J. Sainte-Marie, and I. Vignon-Clementel. *A poroelastic model valid in large strains with applications to perfusion in cardiac modeling.* Computational Mechanics **46** (2010), 91–101.

[6] D. Chapelle and P. Moireau. *General coupling of porous flows and hyperelastic formulations—from thermodynamics principles to energy balance and compatible time schemes.* European Journal of Mechanics-B/Fluids **46** (2014), 82–96.

[7] S. Conte and C. De Boor. *Elementary Numerical Methods.* McGraw-Hill, New York, (1972).

[8] A. Cookson, J. Lee, C. Michler, R. Chabiniok, E. Hyde, D. Nordsletten, M. Sinclair, M. Siebes, and N. Smith. *A novel porous mechanical framework for modelling the interaction between coronary perfusion and myocardial mechanics.* Journal of biomechanics **45** (2012), 850–855.

[9] A. N. Cookson, J. Lee, C. Michler, R. Chabiniok, E. Hyde, D. Nordsletten, and N. Smith. *A spatially-distributed computational model to quantify behaviour of contrast agents in MR perfusion imaging.* Medical image analysis **18** (2014), 1200–1216.

[10] C. Cuenod and D. Balvay. *Perfusion and vascular permeability: basic concepts and measurement in DCE-CT and DCE-MRI.* Diagnostic and interventional imaging **94** (2013), 1187–1204.

[11] T. A. Davis. *Algorithm 832: UMFPACK v4.3—an unsymmetric-pattern multifrontal method.* ACM Transactions on Mathematical Software (TOMS) **30** (2004), 196–199.

[12] S. Di Gregorio, M. Fedele, G. Pontone, A. F. Corno, P. Zunino, C. Vergara, and A. Quarteroni. *A computational model applied to myocardial perfusion in the human heart: From large coronaries to microvasculature.* Journal of Computational Physics **424** (2021), 109836.

[13] P. Eichler, V. Fuka, and R. Fučík. *Cumulant lattice Boltzmann simulations of turbulent flow above rough surfaces.* Computers & Mathematics with Applications **92** (2021), 37–47.

[14] P. Eichler, R. Galabov, R. Fučík, K. Škardová, T. Oberhuber, P. Pauš, J. Tintěra, and R. Chabiniok. *Non-newtonian turbulent flow through aortic phantom: Experimental and computational study using magnetic resonance imaging and lattice boltzmann method.* Computers & Mathematics with Applications **136** (2023), 80–94.

[15] R. Ewing and R. Heinemann. Incorporation of mixed finite element methods in compositional simulation for reduction of numerical dispersion. In 'SPE Reservoir Simulation Symposium'. OnePetro, (1983).

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

[17] R. Fučík, R. Galabov, P. Pauš, P. Eichler, J. Klinkovský, R. Straka, J. Tintěra, and R. Chabiniok. *Investigation of phase-contrast magnetic resonance imaging underestimation of turbulent flow through the aortic valve phantom: Experimental and computational study using lattice boltzmann method.* Magnetic Resonance Materials in Physics, Biology and Medicine **33** (2020), 649–662.

[18] R. Fučík, J. Klinkovský, J. Solovský, T. Oberhuber, and J. Mikyška. *Multidimensional mixed–hybrid finite element method for compositional two-phase flow in heterogeneous porous media and its parallel implementation on GPU.* Computer Physics Communications **238** (2019), 165–180.

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

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

[21] B. Hogan, Z. Shen, H. Zhang, C. Misbah, and A. I. Barakat. *Shear stress in the microvasculature: influence of red blood cell morphology and endothelial wall undulation.* Biomechanics and modeling in mechanobiology **18** (2019), 1095–1109.

[22] H. Hoteit and A. Firoozabadi. *Numerical modeling of two-phase flow in heterogeneous permeable media with different capillarity pressures.* Advances in water resources **31** (2008), 56–73.

[23] M. Jerosch-Herold, R. T. Seethamraju, C. M. Swingen, N. M. Wilke, and A. E. Stillman. *Analysis of myocardial perfusion MRI.* Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **19** (2004), 758–770.

[24] R. Jogiya, S. Kozerke, G. Morton, K. De Silva, S. Redwood, D. Perera, E. Nagel, and S. Plein. *Validation of dynamic 3-dimensional whole heart magnetic resonance myocardial perfusion imaging against fractional flow reserve for the detection of significant coronary artery disease.* Journal of the American College of Cardiology **60** (2012), 756–765.

[25] T. Krüger, H. Kusumaatmaja, A. Kuzmin, O. Shardt, G. Silva, and E. M. Viggen. *The lattice Boltzmann method.* Springer International Publishing **10** (2017), 4–15.

[26] J. Lee, A. Cookson, R. Chabiniok, S. Rivolo, E. Hyde, M. Sinclair, C. Michler, T. Sochi, and N. Smith. *Multiscale modelling of cardiac perfusion.* In 'Modeling the

heart and the circulatory system', Modeling the heart and the circulatory system, Springer (2015), 51–96.

[27] J. Lee, D. Nordsletten, A. Cookson, S. Rivolo, and N. Smith. *In silico coronary wave intensity analysis: application of an integrated one-dimensional and poromechanical model of cardiac perfusion.* Biomechanics and modeling in mechanobiology **15** (2016), 1535–1555.

[28] C. Michler, A. Cookson, R. Chabiniok, E. Hyde, J. Lee, M. Sinclair, T. Sochi, A. Goyal, G. Vigueras, D. Nordsletten, et al. *A computationally efficient framework for the simulation of cardiac perfusion using a multi-compartment darcy porous-media flow model.* International journal for numerical methods in biomedical engineering **29** (2013), 217–232.

[29] S. Plein, S. Ryf, J. Schwitter, A. Radjenovic, P. Boesiger, and S. Kozerke. *Dynamic contrast-enhanced myocardial perfusion MRI accelerated with k-t SENSE.* Magnetic Resonance in Medicine **58** (2007), 777–785.

[30] E. Rohan, J. Turjanicová, and V. Lukeš. *The biot–darcy–brinkman model of flow in deformable double porous media; homogenization and numerical modelling.* Computers & Mathematics with Applications **78** (2019), 3044–3066.

[31] E. C. Sammut, A. D. Villa, G. Di Giovine, L. Dancy, F. Bosio, T. Gibbs, S. Jeyabraba, S. Schwenke, S. E. Williams, M. Marber, K. Alfakih, T. F. Ismail, R. Razavi, and A. Chiribiri. *Prognostic value of quantitative stress perfusion cardiac magnetic resonance.* JACC: Cardiovascular Imaging **11** (2018), 686–694.

[32] U. Schmiedl, M. Ogan, M. Moseley, and R. Brasch. *Comparison of the contrast-enhancing properties of albumin-(gd-DTPA) and gd-DTPA at 2.0 T: and experimental study in rats.* American journal of roentgenology **147** (1986), 1263–1270.

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

[34] C. Sun and L. L. Munn. *Lattice-boltzmann simulation of blood flow in digitized vessel networks.* Computers & Mathematics with Applications **55** (2008), 1594–1600.

[35] N. Westerhof, C. Boer, R. Lamberts, and P. Sipkema. *Cross-talk between cardiac muscle and coronary vasculature.* Physiological Reviews **86** (2006), 1263–1308.

[36] N. Zarinabad, G. L. Hautvast, E. Sammut, A. Arujuna, M. Breeuwer, E. Nagel, and A. Chiribiri. *Effects of tracer arrival time on the accuracy of high-resolution (voxel-wise) myocardial perfusion maps from contrast-enhanced first-pass perfusion magnetic resonance.* IEEE Transactions on Biomedical Engineering **61** (2014), 2499–2506.

Quadratically Integrable Systems with Velocity Dependent Potentials: Generalized Integrals*

Ondřej Kubů
ondrej.kubu@fjfi.cvut.cz

study programme: Mathematical Engineering
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 quadratic integrability of systems with velocity dependent potentials in three-dimensional Euclidean space. Unlike in the case with only scalar potential, quadratic integrability with velocity dependent potentials does not imply separability in the configuration space [1, 2]. The leading order terms in the pairs of commuting integrals can either generalize the forms leading to separation in the absence of a vector potential by adding some terms, or have no relation at all [7] (first noted in [6]). We call such pairs of integrals generalized.

In this workshop contribution we summarize the state of the art concerning these system, with focus on new results [4]. The classes to be considered were classified in [7] and an integrable example in a class not extending any orthogonal coordinate system was presented, later extended in [3]. The article [5] analysed generalized spherical and cylindrical cases. The former did not lead to anything new, the latter includes 3 generalized integrable systems. One of them restricts to the superintegrable helical undulator in infinite solenoid, the only such known system, which has a generalized first order integral.

The new paper [4] focuses on three cases with generalized non-subgroup type integrals, namely elliptic cylindrical, prolate / oblate spheroidal and circular parabolic integrals, together with one case not related to any orthogonal coordinate system. We find two new integrable systems, non-separable in the configuration space, both with generalized elliptic cylindrical integrals. In the other cases, all systems found were already known and possess standard pairs of integrals.

Keywords: integrability, velocity dependent potentials, generalized integrals, non-subgroup integrals, classical mechanics

Abstrakt. Uvažujme kvadratickou integrabilitu na Euklidovském 3D prostoru s potenciálem závislým na rychosti. Na rozdíl od skalárního případu již kvadratická integrabilita neimplikuje separaci v konfiguračním prostru [1, 2]. Členy nejvyššího řádu dvou komutujících kvadratických integrálů mohou tvar odpovídající separaci proměnných v případě bez vectorového potenciálu

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

rozšiřovat o několik členů nebo k němu nemusí mít žádný vztah [7], (poprvé pozorováno v [6]). Takovéto integrály nazýváme zobecněné.

V tomto příspěvku shrnujeme současný stavv poznání o těchto systémech se zaměžením na nové výsledky z [4]. Třídy systémů, které je třeba zkoumat, byly klasifikovány v [7], kde byl také představen systém bez souvislosti se separací proměnných, který byl následně rozšířen v [3]. V článku [5] byly analyzovány případy nazvané zobecněný sférický a cylindrický. První z nich nevedl k ničemu novému, v druhém byly nalezeny tři zobeceněné integrabilní systémy. Po jisté restrikci parametrů se jeden z nich stává superintegrabilní a modeluje šroubovicový undulátor v nekonečném solenoidu. Jedná se o jediný známý superintegrabilní systém se zobecněným integrálem, v tomto případě prvního řádu.

Nový článek [4] se zaměřuje na tři případy zobecňující integrály nepodgrupového typu, konkrétně elipticko cylindrické, protáhlé / zploštělé sferoidální a rotačně parabolické, a navíc jeden případ bez souvislosti s ortogonálními souřadnicemi. Nacházíme dva nové integrabiliní systémy neseparující na konfiguračním prostoru, oba s integrsály zobecněného elipticko-cylindrického typu. V ostatních případech jsou nalezeny pouze již známé nezobecněné systémy.

Klíčová slova: integrabilita, potenciály závislé na rychlosti, zobecněné integrály, integrály nepodgrupového typu, klasická mechanika

Full paper: Hoque, M.F., Kubů, O., Marchesiello, A. et al. *New classes of quadratically integrable systems with velocity dependent potentials: non-subgroup type cases*. Eur. Phys. J. Plus **138** (2023), 845 .

References

[1] V. G. Bagrov, V. N. Shapovalov, and A. G. Meshkov. *Separation of variables in the stationary Schrödinger equation.* Soviet Physics J. **15** (1974), 1115–1119.

[2] S. Benenti, C. Chanu, and G. Rastelli. *Variable separation for natural Hamiltonians with scalar and vector potentials on Riemannian manifolds.* J. Phys. A **42** (2001), 2065–2091.

[3] F. Hoque and L. Šnobl. *Family of nonstandard integrable and superintegrable classical hamiltonian systems in non-vanishing magnetic fields.* J. Phys. A **56** (2023), 165203.

[4] M. F. Hoque, O. Kubů, A. Marchesiello, and L. Šnobl. *New classes of quadratically integrable systems with velocity dependent potentials: non-subgroup type cases*. Eur. Phys. J. Plus **138** (2023).

[5] O. Kubů, A. Marchesiello, and L. Šnobl. *New classes of quadratically integrable systems in magnetic fields: The generalized cylindrical and spherical cases.* Ann. Phys. **451** (2023), 169264.

[6] A. Marchesiello and L. Šnobl. *Superintegrable 3D systems in a magnetic field corresponding to Cartesian separation of variables.* J. Phys. A **50** (2017), 245202, 24.

[7] A. Marchesiello and L. Šnobl. *Pairs of commuting quadratic elements in the universal enveloping algebra of Euclidean algebra and integrals of motion.* J. Phys. A **55** (2022), 145203.

Black Hole Uniqueness in Gravity Conformally Coupled to a Scalar Field *

Tereza Lehečková
lehecter@cvut.cz

study programme: Mathematical Engineering
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Marcello Ortaggio, Department of Algebra, Geometry and Mathematical Physics
Institute od Mathematics, CAS

Josef Schmidt, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This paper presents an introduction to the "no hair" theorems in the form of a review. It first gives a brief overview of the basic ideas and development of theorems (and their possible avoidance) within (electro)vacuum general relativity and different kinds of its generalisations. It then discusses the situation with the scalar field, which under usual conditions does not allow for the appearance of "hair". Finally, it looks in some detail at the results and possible future directions of research in gravity conformally coupled to the scalar field. This theory is interesting both in context of no hair conjecture (remarkable hairy counterexamples have been found) and physical description (e.g. the combination of gravity and other forces).

Keywords: black hole uniqueness, "hairy" black holes, conformally coupled scalar field

Abstrakt. Příspěvek představuje úvod do "no hair"teorémů v podobě review. Nejprve podává krátký přehled základních myšlenek a vývoje teorémů (a jejich možných porušení) v rámci (elektro)vakua v obecné relativitě a různých druhů jeho zobecnění. Následně se věnuje situaci se skalárním polem, které za běžných podmínek výskyt "vlasů"neumožňuje. Konečně se detailněji zaměřuje na výsledky a možné další směry výzkumu v gravitaci konformně spjaté se skalárním polem. Tato teorie je zajímavá jak z hlediska "no hair"doměnky (byly nalezeny pozoruhodné "vlasaté"protipříklady), tak z hlediska fyzikálního popisu (např. kombinace gravitace a ostatních sil).

Klíčová slova: unikátnost černých děr, "vlasaté" černé díry, konformně spjaté skalární pole

1 Introduction

The black hole (BH) uniqueness/no hair theorems (or the overall no hair conjecture expressing the general idea) basically say that black holes in equilibrium can, despite the possibility of their formation in a very complicated process (gravitational collapse) be described by only a few parameters. In the framework of general relativity, these parameters are mass m, angular momentum a and electric (and magnetic, if we allow for

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

the existence of monopoles) charge q and p. Two BHs matching in these parameters are therefore completely identical, unlike other astronomical objects which can be radically different despite agreeing on these values. A nice historically stratified overview is provided by [20]. A detailed description of the basic theorems, proofs and methods can be found in [13].

The first of such theorems was formulated by Israel [14] in 1967, who investigated asymptotically flat vacuum solutions of Einstein's field equations (EFE) with a regular horizon and with conditions to be satisfied by static BHs. The result was that the only possible such spacetime is the Schwarzschild one. He then generalized the theorem to the electrovacuum [15] and revealed the Reissner-Nordström class of spacetimes as the only possibility. Due to the growing body of knowledge about BHs (causal structure, horizon topology...) and new methods (σ models, harmonic mapping...), the assumptions of the theorem and its proofs were subsequently varied many times.

Although the question of the uniqueness of Kerr's BH among stationary spacetimes was already raised in [14], it waited for confirmation until 1975, when Robinson proved it [21]. Thus, the overall result in GR is that the most general electrovacuum BH in an asymptotically flat spacetime with a regular horizon is Kerr-Newman. Further investigations and generalizations involved more BHs, other dimensions, and of course the addition of different matter fields or cosmological constant [13].

The aforementioned generalizations have made the subject considerably more complicated. It turned out that many of the assumptions and methods valid in GR cease to work with some fields (and/or cosmological constants) and the way of study must therefore be very different. A number of peculiar "hairy" black holes have been found (e.g. in Einstein-Yang-Mills theories [22]), but often problematic due to various divergences or instability of solutions.

Current research includes proofs of no hair theorems, search for counterexamples (analytical and numerical), their physical analysis (e.g. thermodynamics) or numerical stability tests (e.g. [9]). In this paper we give an example of GR theorems and their background, followed by a generalization to scalar fields. Finally, we will focus on one particular interesting theory - gravity conformally coupled to the scalar field. In it, we will look at the solutions and theorems found and discuss possibilities for further research.

2 Black hole uniqueness in GR

Before we get to the particular theorems, let us outline the framework in which we will work [20]. First, what is a **BH spacetime** - it is a spacetime having an **event horizon** $\mathcal{H}$. This consists of the **future event horizon** $\mathcal{H}^+$(i.e. boundary of the set of events in the causal past of future null infinity) and possibly the **past event horizon** $\mathcal{H}^-$ (if time orientation changed). The region hidden below the horizon is referred to as the **BH** itself. No hair theorems always concern the horizon and its exterior, the so-called **domain of outer communication** $<< M >>$ (i.e. set of events from which there exist both future and past directed curves extending to arbitrary large asymptotic distances), not the interior.

We also emphasize that the theorems only concern BH, not naked singularities. The parameters used to describe them here are easily defined by the ADM formalism due to

the high symmetries and asymptotic flatness. In order to make the solutions physical, some version of the energy condition is usually assumed [11], e.g. a dominant energy condition ($T^{ab}W_aW_b \geq 0$; $T^{ab}W_a$ non-space-like for every time-like W^a). We now start with Israel's theorem in a modern Robinson's version and directly discuss its assumptions folowing [20].

Theorem: The most general static, asymptotically flat single BH solution of EFE with regular horizon is Schwarzschild solution.

Static: $\exists$ time-like hypersurface orthogonal Killing vector field (KVF), i.e. $k^\alpha k_\alpha < 0$; $k_{[\alpha}\nabla_\beta k_{\gamma]} = 0$ and adapted coordinate system $(t, \boldsymbol{x})$ in which $k^\alpha = (1, \boldsymbol{0})$ and $ds^2 = -v^2dt^2 + g_{ab}dx^adx^b$ (where v and g_{ab} are independent on t).

Asymptotically flat: $g_{ab} = (1 + 2mr^{-1})\delta_{ab} + h_{ab}$; ; $v = 1 - mr^{-1} + \mu$; $m = \text{const}$, where x^a are asymptotically Euclidian coordinates and h_{ab}, μ are all $O(r^{-2})$ and $O(r^{-3})$ in 1st derivatives as $r = (\delta_{ab}x^ax^b) \to \infty$.

Single regular horizon: $B := \mathcal{H}^+ \cap \mathcal{H}^-$ is regular compact, connected boundary to Σ as $v \to 0$, where Σ are regular hypersurfaces $t = \text{const}$, $0 < v < 1$.

EFE: in the adapted coordinates given above, they read $R_{tt} \equiv vD^aD_av = 0$, $R_{ta} \equiv 0$, $R_{ab} \equiv R_{ab} - v^{-1}D_aD_bv = 0$.

We do not have space to prove the theorem (although it is not very long [20]) here, but we will try to outline its basic ideas. It uses the function $w := -\frac{1}{2}\nabla_{[\alpha}k_{\beta]}\nabla^\alpha k^\beta = g^{ab}v_{,a}v_{,b}$ (which, among other things, can be shown to be constant on B and equal to the square of surface gravity) and the EFE expressed in terms of w and v. By integrating over Σ, using boundary conditions, the Gauss-Bonnet theorem and assumptions, we obtain a set of observations, identities and inequalities that give us information about the spacetime. These are so restrictive that with them we can show that the so-called Cotton tensor R_{abc} is zero (and hence g_{ab} is conformally flat) and determine the particular shape of w. This is already enough to reveal the Schwarzschild in the assumed metric.

Concepts from assumptions can be defined in multiple equivalent ways. Israel's original [14] assumptions looked very different and were debated. Today we know they were stronger than necessary. In addition to staticity, there were three conditions expressed by Σ guaranteeing asymptotic flatness and geometric regularity. Furthermore, there was an assumption/condition key in the proof that basically enforced the spherical topology of equipotential surfaces $v = \text{const}$ in Σ and also the possibility of covering Σ with one coordinate system (with v as one of the coordinates). Using this system, it was then possible to construct a number of identities that already restricted possible solutions to Schwarzschild.

In the extension to electrovacuum we add an electromagnetic field to $<< M >>$, i.e. Einstein-Hamilton action will be $S = \frac{1}{4\pi}\int d^4\sqrt{-g}\left(\frac{R}{4} - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}\right)$ (using units $G = c = 4\pi\epsilon_0 = 1$). The modification of the proof of [15] here was essentially to show that the equipotential surfaces of this field coincide with the (surface) gravitational ones and,

finally, that they are necessarily spherical.

A generalization to stationary spacetimes (time-like KVF k^α need not be hypersurface orthogonal), i.e. a proof of the uniqueness of the Kerr (and with the addition of charge later Kerr-Newman) solution, appeared in 1975 [21]. It exploited the new insights of Hawking and Ellis [11] that $<< M >>$ of a stationary BH must be axisymmetric and its topology is $S^2 \otimes R$. The proof used Carter's [8] method of converting the EFE to a boundary value problem of a system of elliptic partial differential equations on a two-dimensional manifold.

3 Scalar field

The scalar field has a specific position in physics. It is, from a certain point of view, the simplest "matter" that can be added to the theory we want to investigate. It is both a toy model but also a (potentially) real thing (Higgs field, physics behind the standard model, models of dark matter/energy...)

It is also special in terms of no hair theorems. Naturally, scalar fields were among the first generalizations of the theorems to be tested. While hairy black holes were found for many ther kinds of fields [22] already in the 1980s, the no hair conjecture still seemed to hold for scalar fields. A number of no hair theorems (e.g. [2, 4]) have been proved which rule out scalar hair. In this section we look at look at an example of these theorems and some issues related to it. We first formulate a generalization of Bekenstein's no hair theorem, restated by [12], and explane its assumptions.

Theorem: There is no regular rotating, stationary, asymptotically flat BH spacetime with scalar hair for which: scalar field Φ is minimally coupled, inherits the spacetime symmetries and its potential V obeys $\Phi V' \geq 0$ everywhere with $\Phi V' = 0$ for (at most) some dicrete values Φ_i.

Canonical minimal coupling: means basically no mixed terms containing R and Φ at the same time. Thus, the action is of the form $S = \frac{1}{4\pi} \int d^4x \sqrt{-g} \left(\frac{R}{4} - \frac{1}{2} \bigtriangledown_\mu \Phi \bigtriangledown^\mu \Phi - V(\Phi) \right)$ and scalar field equations (SFE) are $\bigtriangledown_\mu \bigtriangledown^\mu \Phi - V'(\Phi) = 0$.

Symmetry inheritance: in adapted coordinates (t, r, θ, ϕ) stationar spacetime has two KVF ∂_t and ∂_ϕ, so $\partial_t \Phi = \partial_\phi \Phi = 0$.

Potential condition: is in fact energy condition. It restricts the (physical) potentials to which we can apply the theorem, the version mentioned here being the simplest. To make the solution physical, we only need to assume a weak energy condition ($T^{ab} W_a W_b \geq 0$ for every time-like W_a), the results are the same, the proofs are different.

The proof starts by multiplying SFE by Φ and integrating by parts over the exterior of BH, i.e. $\int d^4x \sqrt{-g} \left(\Phi \bigtriangledown_\mu \Phi \bigtriangledown^\mu \Phi + \Phi V' \right) = 0$. There are two boundary terms - $\mathcal{H}$ and infinity. The latter vanishes due to asymptotic flatness and the former due to field symmetries ($\mathcal{H}$ is also a Killing horizon). We therefore get $\int d^4x \sqrt{-g} \left(\bigtriangledown_\mu \Phi \bigtriangledown^\mu \Phi + \Phi V' \right) = 0$. Now both terms of the integrand are non-negative (the first its orthogonality to both

KV, the second from the potential condition). Therefore, for equality to hold, it must be $\Phi = 0$ (if $\Phi = \Phi_i$ it is actually a cosmological constant and asymptotic flatness no longer holds).

A few remarks on the theorem: note that EFEs do not appear anywhere in the proof. This is quite typical for no hair theorems with added field, they usually work mainly with field equations and boundary conditions. Next, what does the word regular mean here: in the context of (potentially) hairy black holes with a field, it means regularity of both the gravitational and other fields (meaning $T^{ab}T_{ab}$) on and outside the horizon. (We will return to this in the next section.)

We notice that by violating some of the assumptions "hairy" BHs can be constructed [23] (however often unphysical). We will be most interested in the first assumption. This can of course be violated in various ways (e.g. higher order gravity theories), but in the context of the next section let us focus on non minimal coupling. That is, adding a mixed term (gravity and scalar field) to the action, and hence a term with curvature to the SFE. Thus, the action to be addressed is

$$S = \frac{1}{4\pi}\int d^4x\sqrt{-g}\left(\frac{R}{4} - \frac{1}{2}\bigtriangledown_\mu \Phi \bigtriangledown^\mu \Phi - V(\Phi) - \frac{1}{2}\xi R\Phi^2\right),$$

where ξ is so called **coupling constant**. There are no hair theorems about this general theory as well, e.g. [5], but we will specify them more in the next section.

4 Conformal coupling

In this section we focus on the gravity conformally coupled to the scalar field, i.e. the choice of coupling constant $\xi = \frac{1}{6}$. In the simplest version (**conformal scalar vacuum**), the action we are investigating takes the form

$$S = \frac{1}{4\pi}\int d^4x\sqrt{-g}\left(\frac{R}{4} - \frac{1}{2}\bigtriangledown_\mu \Phi \bigtriangledown^\mu \Phi - \frac{1}{12}R\Phi^2\right)$$

and SFE read $\bigtriangledown_\mu \bigtriangledown^\mu \Phi - \Phi\frac{R}{6} = 0$. It is also possible to add Λ or $V(\Phi)$ to the action.

One of the reasons why this theory is interesting is that when performing the **conformal transformation** $g_{\mu\nu} \to \tilde{g}_{\mu\nu} = \Omega^2 g_{\mu\nu}$ and $\Phi \to \tilde{\Phi} = \Phi/\Omega$, the SFEs are invariant. (EFEs - the same as in GR - are also invariant, but the overall action is not). Conformal transformation is crucial not only in GR (e.g. it is used to investigate the causal structure of spacetimes or to classify them) but also in the context of efforts to unify the fundamental forces (conformal invariance is also expected from quantum models).

The first and still discussed solution of the above equations is the so-called **BBMB solution** found independently by a group of Russian authors in 1970 [7] and by Bekenstein in 1973 [2]. In Bekenstein's case it was also the discovery of a method to generate (under certain conditions) from a minimally coupled scalar field solution a conformally coupled one. It is of the form

$$ds^2 = -\left(1 - \frac{M}{r}\right)^2 dt^2 + \left(1 - \frac{M}{r}\right)^{-2} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2), \quad \Phi = \frac{\sqrt{3}M}{r - M}.$$

This solution is essentially hairy BH but there are a few issues to adress. The field is clearly diverging on the horizon $r = M$. This initially seemed to be a reason not to accept it as "physically allowed". However, since this divergence does not translate into a metric and a test scalar particle passing through the horizon should not experience anything special [3], this does not seem to be such a problem. The BBMB solution is one-parametric, geometrically identical to the extreme Reissner-Nordström for $\mid Q_E \mid = M$, so we have no new parameter/scalar charge. Such a case is called **secondary hair** (the case where a new parameter would appear is **primary hair**). The solution is unstable, so there is after all a kind of physical problem. A very important fact is that the solution is unique. In 1991 it was proved [24] to be the only non-trivial static asymptotically flat solution in gravity conformally coupled to a scalar field.

At this point, we make a generalization concerning Λ i.e. the geometry is no longer asymptotically flat, but asymptotically (anti-)de Sitter. The cosmological constant is both a natural generalization and a possible step towards a realistic model of our universe. When added to pure GR, (albeit with more complicated proofs) the no hair theorem can be extended to this case [10]. In the case of a scalar field, however, the situation changes considerably. For positive Λ the no hair theorems can be extended, but for negative ones it can be shown that solutions exist. (This applies to both minimal and conformal coupling) [23]. It is also worth mentioning a number of interesting results concerning Λ in 3D (e.g. [25]).

The reason of this "theorem transfer" is the relation of minimal and conformal coupling via conformal transformation. When we apply it to a conformally coupled scalar field, we get an action in which quantities from the transformed metric and a new, minimally coupled field stand out, namely $\varphi = \sqrt{6}\tanh^{-1}\frac{\Phi}{\sqrt{6}}$ (a similar transformation can be done for a general coupling constant, but not with a general manageable integral). The transformed potential may be unphysical, even if the original $V(\Phi)$ is not, which is however not a problem. For $\Lambda \neq 0$, the potential also arises for $V(\Phi) = 0$, specifically $U(\varphi) = \frac{\Lambda\Phi^2(12-\Phi^2)}{(6-\Phi^2)^2}$.

The advantage of this method is that the obtained minimally coupled systems are generally easier treatable (both numerically and analytically) and generally better studied. The disadvantage, on the other hand, is that the conformal transformation is not always applicable (if $\Omega = 1 - \frac{\Phi}{6} = 0$ is on the horizon or outside, the transformation is invalid). This method can also be used to establish no hair theorems by using their analogues for minimal coupling [23]. For example, it can be seen that for a potential-free (i.e., massless) field or a mass field of type $V(\Phi) = \frac{\mu^2}{2}\Phi^2$ and $\Lambda > 0$, the transformed potential is convex and thus satisfies one of the Bekenstein variations of the energy/potential condition [12]. From this we know that there are no regular solutions (so the most general one is Schwarzschild-de-Sitter BH). On the contrary, for $\Lambda < 0$ well behaved solutions are automatically offered as long as the field mass is not too large [23].

As for the solutions that have been found within $\Lambda < 0$ this has often been done by numerical methods and the results cannot always be expressed explicitly. Nevertheless, solutions have been successfully analyzed, for example [23], which are even stable against perturbations.

For $\Lambda > 0$ there is a so-called **MTZ solution** from 2003, [17], which is essentially a generalization of the BBMB solution with quartic potential (i.e. the above mentioned

theorem does not apply to it). Its form is

$$ds^2 = -\left[-\frac{\Lambda}{3}r^2 + \left(1 - \frac{GM}{r}\right)^2\right] dt^2 + \left[-\frac{\Lambda}{3}r^2 + \left(1 - \frac{GM}{r}\right)^2\right]^{-1} dr^2 + r^2 d\Omega^2,$$

$$\Phi(r) = \sqrt{\frac{3}{4\pi}} \frac{\sqrt{G}M}{r - GM}.$$

As can be seen, this field does not diverge at the event horizon (which is no longer extreme). The geometry is identical to Reissner-Nordstrom-de Sitter BH with $\mid Q_E \mid = M$, so there are three horizons (inner, event and cosmological). The solution has been extensively analyzed and discussed. It has been found to be unstable and there even seem to be no stable solutions for the given conditions [9].

5 Conclusion and possible extensions

After a brief historical summary and outline of the no hair theorems in the first section, we introduced (in the second section) the Israel theorem (restated by Robinson), outlined its proofs and discussed its assumptions. We continued with information on its direct generalization to electrovacuum and stationary spacetimes and mentioned other extensions and generalizations in the problem (Λ, fields, other theories and dimensions...). In the third section we focused specifically on scalar fields. Again, we gave an example of the no hair theorem and the idea of its proof. We discussed the special features of the scalar field and the importance of coupling. In the forth section we focused exclusively on the conformally coupled scalar field, because of its specificity both in terms of physics and solvability. Here we discussed the conformal transformation (its meaning and applications) and presented some solutions - BH with scalar hair, e.g. BBMB and MTZ and discussed their properties. We also returned in detail to the generalization containing Λ and its relevance in terms of no hair theorems in (not only) conformal coupling.

As far as other research directions are concerned, the possibilities of extending the no hair theorems from GR are abundant. However, if we focus only on gravity conformally coupled to scalar fields, there are many possibilities here as well. Traditionally, these are variations of the assumptions, methods and proofs in no hair theorems and the search for new ones. Then, of course, there is the search for new solutions of hairy black holes and the analysis of these solutions in terms of both their physicality (stability, possible divergences) and physical properties (e.g. thermodynamic, the appearance of so-called spontaneous dressing up [18]).

Finally, it is also possible to take gravity conformally coupled to scalar field simply as a new theory and treat it like GR, i.e. to make some classification of spacetimes, BH spacetimes and their typical/possible properties. Another, somewhat broader possibility is to add to this theory other fields sharing conformal invariance and study no hair theorems and hairy BHs in them. The next step could then be to modify the gravity part of the theory itself and study how things change in e.g. higher order gravity theories.

Although many hairy BHs are known today, it is still true that BHs are generally described by a very small number of parameters. Putting this in a broader context, BH

uniqueness is related to the information paradox (as e.g. a collapsing star loses its hair and where/if the information is then stored somewhere), thermodynamic analysis (stability and behaviour towards disturbances of equilibrium, new thermodynamic properties of hairy BHs), or the cosmic censorship hypothesis (originally the no hair theorems were considered purely in the context of gravitational collapse results, although this is no longer the case, many of them seem to support the hypothesis).

References

[1] J. D. Bekenstein. *Nonexistence of Baryon Number for Static Black Holes.* Phys. Rev. D 5, 1239 (1972)

[2] J. D. Bekenstein. *Exact solutions of Einstein-Conformal scalar equations.* Ann. Phys. 82 pp. 535-547 (1974)

[3] J. D. Bekenstein. *Black holes with scalar charge.* Ann. Phys. 91, pp. 75-82 (1975)

[4] J. D. Bekenstein. *Novel "no-scalar-hair" theorem for black holes.* Phys. Rev. D 51, R6608(R) (1995)

[5] J. D. Bekenstein and A. Mayo. *No hair for spherical black holes: Charged and non-minimally coupled scalar field with self-interaction.* Phys. Rev. D 54(8) pp. 5059-5069 (1996)

[6] J. D. Bekenstein. *Black hole hair: twenty-five years after.* Physics. Proceedings, 2nd International A.D. Sakharov Conference, Moscow, Russia, pp. 216-219 (1996)

[7] N. M. Bocharova, K. A. Bronnikov, and V. N. Melnikov. *On an accurate solution of the system of equations of Einstein and of a scalar field empty of mass.* Vestn. Mosk. Univ. Ser. III Fiz. Astron., (6) pp. 706 (1970)

[8] B. Carter. *Axisymmetric black hole has only two degrees of freedom.* Phys. Rev. Lett. 26 pp. 331-333 (1971)

[9] G. Dotti, R. J. Gleiser and C. Martinéz. *Static black hole solutions with a self interacting conformally coupled scalar field.* Phys. Rev. D 77 104035 (2008)

[10] G. Gotz. *On the cosmological "no-hair" conjecture.* Phys. Lett. A 128, pp. 129 – 132 (1988)

[11] S. W. Hawking and G. F. R. Ellis. *The Large Scale Structure of Space-Time.* Cambridge University Press (1973)

[12] C. A. R. Herdeiro and E. Radu. *Asymptotically flat black holes with scalar hair: a review.* Int. J. Mod. Phys. D 24 09, 1542014 (2015)

[13] M. Heusler. *Black hole uniqueness theorems.* Cambridge University Press (1996)

[14] W. Israel. *Event horizons in static vacuum spacetimes.* Phys. Rev. 164, 1776 (1967)

[15] W. Israel. *Event horizons in static electrovac spacetimes*. Commun. Math. Phys. 8, 245 (1968)

[16] C. Martínez. *Black holes with a conformally coupled scalar field*. Part of Quantum Mechanics of Fundamental Systems: The Quest for Beauty and Simplicity : Claudio Bunster Festschrift, pp. 167-180 (2009)

[17] C. Martinéz, R. Troncoso and J. Zanelli. *de Sitter black hole with conformally coupled scalar field in four dimensions.* Phys. Rev. D 67 024008 (2003)

[18] C. Martinéz, R. Troncoso and J. Zanelli. *Exact black hole solution with minimally coupled scalar field.* Phys. Rev. D 70 084035 (2004)

[19] E. Radu. *Conformally coupled scalar solitons and black holes with negative cosmological constant*. Phys. Rev. D 72, 024017 (2005)

[20] D. C. Robinson. *Four decades of black hole uniqueness theorems*. Part of The Kerr Spacetime: Rotating Black Holes in General Relativity, pp. 115-143 (2009)

[21] D. C. Robinson. *Uniqueness of Kerr black hole.* Phys. Rev. Lett. 34, 905 (1975)

[22] M. S. Volkov and D. V. Gal'tsov. *Gravitating non-abelian solutions and black holes with Yang-Mills fields*. Phys. Rep. 319 (1-2), pp. 1-83 (1999)

[23] E. Winstanley. *On existence of conformally coupled scalar field hair for black holes in (anti-)de Sitter space*. Found. Phys. 33, pp. 111–143 (2003)

[24] B. C. Xanthopoulos and T. Zannias. *The uniqueness of the Bekenstein black hole.* J. Math. Phys. 32, pp. 1875–1880 (1991)

[25] M. Banados, C. Teitelboim and J. Zanelli. *The Black hole in three-dimensional spacetime.* Phys. Rev. Lett. 69, pp. 1849-1851 (1992)

Dumont-Thomas Numeration Systems for $\mathbb{Z}$*

Jana Lepšová
lepsojan@fjfi.cvut.cz

study programme: Mathematical Engineering
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, LaBRI
University of Bordeaux, CNRS, Bordeaux INP

Abstract. Numeration systems for representing nonnegative integers were proposed in [1]. They are based on right-infinite fixed points of substitutions and we refer to them as to the Dumont–Thomas numeration systems. We extend the Dumont–Thomas numeration systems to $\mathbb{Z}$ by considering two-sided periodic points of substitutions. This allows us to represent any integer in $\mathbb{Z}$ by a finite word (starting with `0` when nonnegative and with `1` when negative). We show that a certain automaton naturally associated with a given substitution returns the letter at position $n \in \mathbb{Z}$ of the corresponding periodic point when fed with the representation of n. The Dumont–Thomas numeration systems can be naturally extended to $\mathbb{Z}^d$, for every $d \geq 2$. We give an equivalent characterization of the numeration systems in terms of a total order on a regular language. Lastly, using particular periodic points of substitutions, we recover the well-known two's complement numeration system [2, §4.1] and the Fibonacci analogue of the two's complement numeration system [3], which can be used to describe a particular Wang tiling [4].

Keywords: substitution, numeration system, automaton, two's complement

Abstrakt. Numerační systémy pro reprezentaci nezáporných celých čísel byly definovány v [1]. Jejich definice je založena na pevných bodech substitucí a podle jmen autorů se nazývají Dumont–Thomas numerační systémy. V tomto příspěvku rozšiřujeme Dumont–Thomas numerační systémy na $\mathbb{Z}$ pomocí oboustranných periodických bodů substitucí. Díky tomu můžeme reprezentovat jakékoli celé číslo konečným slovem (které začíná symbolem `0`, pokud je číslo nezáporné, a které začíná symbolem `1`, pokud je číslo záporné). Dokážeme, že určitý automat, který je přirozeně svázán s danou substitucí, vypíše písmeno na pozici $n \in \mathbb{Z}$ daného periodického bodu, pokud na vstupu zadáme reprezentaci n. Dumont–Thomas numerační systémy mohou být rozšířeny na $\mathbb{Z}^d$, pro každé $d \geq 2$. Dále tyto numerační systémy charakterizujeme pomocí určitého úplného uspořádání regulárního jazyka. Nakonec ukážeme, že příkladem Dumont–Thomas numeračních systémů pro $\mathbb{Z}$ je známý dvojkový doplněk [2, §4.1] a také Fibonacciho obdoba dvojkového doplňku [3], kterou lze použít k popisu určitého Wangova dláždění [4].

*This work has been supported by The French Institute in Prague and the Czech Ministry of Education, Youth and Sports through the Barrande fellowship programme, Agence Nationale de la Recherche through the project Codys (ANR-18-CE40-0007), and the support by Grant Agency of Czech Technical University in Prague, through the project SGS23/187/OHK4/3T/14.

Klíčová slova: substituce, numerační systém, automat, dvojkový doplněk

Full paper: S. Labbé, J. Lepšová. *Dumont-Thomas numeration systems for* $\mathbb{Z}$. Under review in Integers. Preprint accessible at `https://arxiv.org/abs/2302.14481`.

References

[1] J.-M. Dumont, A. Thomas. *Systemes de numeration et fonctions fractales relatifs aux substitutions*. Theoretical Computer Science **65** (1989), 153—169.

[2] D.E. Knuth. *The art of computer programming. Vol. 2: Seminumerical algorithms*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont, (1969).

[3] S. Labbé, J. Lepšová. *A Fibonacci's complement numeration system*. To appear in RAIRO - Theoretical Informatics and Applications (2023). Preprint accessible at `https://arxiv.org/abs/2205.02574`.

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

Palatini Variation in Generalized Geometry and String Effective Actions*

Filip Moučka
mouckfil@cvut.cz

study programme: Mathematical Engineering
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Branislav Jurčo, Mathematical Institute
Faculty of Mathematics and Physics, Charles University

Jan Vysoký, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Courant algebroids are vector bundles with a certain additional structure. This structure allows to introduce a so-called generalized Riemannian geometry of Courant algebroids. In particular, there are generalizations of the notions of metric, connection, torsion, Levi-Civita connection, and curvature. In paper [1], it was shown that the type IIB supergravity equations can be very elegantly formulated in terms of generalized Riemannian geometry. In the follow up paper [2] the result was further extended for heterotic supergravity.

In our paper, we develop the Palatini formalism within the framework of generalized Riemannian geometry of Courant algebroids. In this context, the Palatini variation of a generalized Einstein–Hilbert–Palatini action - formed using a generalized metric, a Courant algebroid connection (in contrary to the ordinary case, not necessarily a torsionless one) and a volume form - leads naturally to a proper notion of a generalized Levi-Civita connection and low-energy effective actions of string theory.

Keywords: Generalized geometry, Courant algebroid connection, generalized torsion and curvature, Einstein-Hilbert action, Palatini formalism, supergravity

Abstrakt. Courantovy algebroidy jsou vektorové bandly s jistou dodatečnou strukturou. Tato struktura umožňuje na Courantových algebroidech vybudovat takzvanou zobecněnou Riemannovu geometrii. Konkrétně lze zavést zobecnění pojmů metrika, konexe, torze, Levi-Civitova konexe a křivost. V článku [1] autoři ukázali, že rovnice supergravitace typu IIB lze velmi elegantně formulovat v řeči zobecněné Riemannovy geometrie. V navazujícím článku [2] byl tento výsledek rozšířen pro heterotickou supergravitaci.

V našem článku jsme vytvořili Palatiniho formalismus požívajíce jazyk zobecněné Riemannovy geometrie na Courantových algebroidech. V tomto případě Palatiniho variace zobecněné Einstein-Hilbert-Palatiniho akce, jejímiž dynamickými proměnnými jsou zobecněná metrika, konexe na Courantově algebroidu (na rozdíl od klasického případu ne nutně bez torzní) a forma objemu, vede přirozeně k vlastními pojmu zobecněné Levi-Civitovy konexe a nízko energetickým efektivním akcím strunové teorie.

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

Klíčová slova: Zobecněná geometrie, konexe na Courantových algebroidech, zobecněná torze a křivost, Einstein-Hilbertova akce, Palatiniho formalismus, supergravitace

Full paper: Branislav Jurčo; Filip Moučka; Jan Vysoký. Palatini variation in generalized geometry and string effective actions. *Journal of Geometry and Physics*, **191**, 2023.

References

[1] B. Jurčo; J.Vysoký. Courant algebroid connections and string effective actions. *Proceedings of Tohoku Forum for Creativity, Special volume: Noncommutative Geometry and Physics IV*, 2016.

[2] J. Vysoký. Kaluza-Klein reduction of low-energy effective actions: geometric approach. *Journal of High Energy Physics*, **143**, 2017.

Dirac Operator on Star-Shaped Graphs

Václav Růžek
ruzekva2@fjfi.cvut.cz

study programme: Mathematical Engineering
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Matěj Tušek, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Dirac operator on star-shaped graphs is introduced. All self-adjoint realisations are given by transmission condition at the vertex in formalism of boundary triples. From the boundary triple Krein's resolvent formula can be quickly deduced. Finally, the correspondence with relativistic point interaction in one dimension is demonstrated.

Keywords: Dirac operator, relativistic quantum graphs

Abstrakt. Nejprve zavedeme Diracův operátor na hvězdicovitém grafu. Všechny jeho samosdružené realizace jsou popsány přechodovými podmínkami v jeho vrcholu. K popisu těchto podmínek použijeme formalismus tzv. *boundary triples*. Díky tomuto formalismu můžeme rychle napsat též Kreinovu formuli. Na závěr je demonstrováno, jak jeden jednoduchý graf přesně odpovídá zavedení relativistické bodové interakce na reálné přímce.

Klíčová slova: Diracův operátor, relativistické kvantové grafy

1 Introduction

The aim of this brief paper is to introduce a star-shaped graph as a configuration space for relativistic quantum mechanics of massless half-integer spin particles. These are subject to Dirac equation, hence the Dirac operators will be studied on this graph. The required definitions and notation is introduced in section 2.

The behaviour is assumed to be free on the edges with some interactions at the vertices. The nature of interactions is closely related to transmission conditions at the vertices. The theory of boundary triples (see [1]) will be used to efficiently describe suitable transmission conditions resulting in self-adjoint operators in section 3.

Next, the section 4 provides the Krein resolvent formula. The most important result which relates the spectrum and resolvents between two self-adjoint Dirac operators.

Finally, the section 5 discusses the special case of the graph with two edges. This case can be directly related to the relativistic point interaction in one dimension studied in [3].

2 Maximal Dirac operator on star-shaped graph

Firstly, let us familiarise ourselves with the configuration space.

A star-shaped graph is a graph Γ with a single vertex $V = \{v\}$ and arbitrary finite number of external edges $E = \{e_i | i \in \{1, \dots, n\}\}$ (no internal edges $I = \emptyset$). The boundary map ∂ assigns the single vertex to every external edge $\forall i \in \{1, \dots, n\} : \partial(e_i) = v$.

On the star-shaped graph the following Hilbert space can be introduced

$$\mathscr{H} = \bigoplus_{i=1}^{n} L^2(\mathbb{R}^+; \mathbb{C}^2) \tag{1}$$

which is a direct sum of n copies of the Hilbert space of $\mathbb{C}^2$-valued square-integrable functions on half-line, each copy associated with one external edge. Elements Ψ of the $\mathscr{H}$ can be described by $2n$-tuple of $L^2(\mathbb{R}^+; \mathbb{C})$ functions

$$\Psi = \begin{pmatrix} \psi_1^1 & \psi_2^1 & \cdots & \psi_n^1 & \psi_1^2 & \cdots & \psi_n^2 \end{pmatrix}^T \tag{2}$$

where for each i the pair $\begin{pmatrix} \psi_i^1 & \psi_i^2 \end{pmatrix}^T$, frequently abbreviated as ψ_i, represent function from $L^2(\mathbb{R}^+; \mathbb{C}^2)$ associated with the edge e_i. In other words, in ψ_i^j the index i is the edge index and the index j is the spinor index while in Ψ the first spinor components of all edges are written first followed by all the second spinor components. To form a Hilbert space, the direct sum (1) is endowed with a scalar product linear in the second argument.

$$\langle \Psi | \Phi \rangle_{\mathscr{H}} = \sum_{i=1}^{n} \int_{\mathbb{R}^+} \left(\overline{\psi_i^1(x_i)} \varphi_i^1(x_i) + \overline{\psi_i^2(x_i)} \varphi_i^2(x_i) \right) \mathrm{d}x_i$$

After the configuration space (the graph) and the quantum state space (the Hilbert space), relativistic quantum mechanics for massless half-integer spin particles will be considered, hence the role of a Hamiltonian will be played by a Dirac operator.

On the space $\mathscr{H}$ we define a differential operator

$$D_{max} = \bigoplus_{i=1}^{n} -\mathrm{i} d_{x_i} \otimes \sigma_1 \tag{3}$$

acting on the direct sum of Sobolev spaces $\operatorname{Dom} D_{max} = \bigoplus_{i=1}^{n} H^1(\mathbb{R}^+; \mathbb{C}^2)$. This is the maximal domain where the differential expression makes sense. The operator D_{max} is called the maximal Dirac operator.

With the same expression as in (3) we define another Dirac operator D_0 on the domain $\bigoplus_{i=1}^{n} H_0^1(\mathbb{R}^+; \mathbb{C}^2)$ where H_0^1 stands for the Sobolev space with zero-trace functions. D_0 is a densely defined closed symmetric operator and its adjoint is $D_0^* = D_{max}$ and thus D_{max} is closed as well.

3 Boundary triple for D_{max}

The Dirac operators D_0 and D_{max}, defined in section 2, are not self-adjoint and therefore they are not proper quantum observables. We intend to study self-adjoint restrictions of D_{max}, which are at the same time self-adjoint extensions of D_0, and thus enabling

the quantum interpretation as Hamiltonian. To describe all self-adjoint extensions, the theory of boundary triples will be used with [1] as a main reference.

Define $\mathscr{G} = \mathbb{C}^n$ as an auxiliary Hilbert space with the standard scalar product $\langle\cdot|\cdot\rangle_{\mathscr{G}}$ and define a rectangular matrix $G_{a,b} \in \mathbb{C}^{n,2n}$ with parameters $a, b \in \mathbb{C}$ as a block matrix

$$G_{a,b} = \begin{pmatrix} aI_n & bI_n \end{pmatrix} \tag{4}$$

where I_n is the identity matrix in $\mathbb{C}^{n,n}$.

Recall that for every $\psi \in H^1(\mathbb{R}^+, \mathbb{C}^2)$ the boundary value $\psi(0)$ is well defined in the sense of the trace operator. By $\Psi(0)$ for $\Psi \in \operatorname{Dom} D_{max}$ is meant $\begin{pmatrix}\psi_1^1(0) & \psi_2^1(0) & \cdots & \psi_n^1(0) & \psi_1^2(0) & \cdots & \psi_n^2(0)\end{pmatrix}^T \in \mathbb{C}^{2n}$.

Define two linear maps $\Gamma_1, \Gamma_2 : \operatorname{Dom} D_{max} \to \mathscr{G}$ as

$$\begin{aligned} \forall \Psi \in \operatorname{Dom} D_{max} : \qquad & \Gamma_1 \Psi = G_{1,0}\Psi(0), \\ & \Gamma_2 \Psi = G_{0,\mathrm{i}}\Psi(0). \end{aligned}$$

On the right-hand side there is a matrix multiplication of n by $2n$ matrix with $2n$ column vector, hence the result is indeed in $\mathscr{G} = \mathbb{C}^n$. Essentially, Γ_1 picks out the traces of the first spinor components on all edges and Γ_2 picks out the traces of the second spinor components on all edges multiplied by i.

Proposition 1. The triple $(\mathscr{G}, \Gamma_1, \Gamma_2)$ defined above is a boundary triple for D_{max} as introduced in [1, Definition 1.7].

Proof. The first condition to verify is that

$$\langle \Psi | D_{max}\Phi\rangle_{\mathscr{H}} - \langle D_{max}\Psi|\Phi\rangle_{\mathscr{H}} = \langle \Gamma_1\Psi|\Gamma_2\Phi\rangle_{\mathscr{G}} - \langle \Gamma_2\Psi|\Gamma_1\Phi\rangle_{\mathscr{G}}$$

for every $\Psi, \Phi \in \operatorname{Dom} D_{max}$. Note that on the left-hand side is the scalar product on $\mathscr{H}$ and on the right-hand side is the scalar product on $\mathscr{G}$. From now on, we will not differentiate the products with index because it should be evident which one is intended. The equality can be seen by a direct computation.

$$\langle D_{max}\Psi|\Phi\rangle = \sum_{i=1}^{n} \langle(-\mathrm{i}d_{x_i} \otimes \sigma_1)\psi_i|\varphi_i\rangle = \sum_{i=1}^{n}\sum_{j=1}^{2} \int_0^{+\infty} \overline{-\mathrm{i}d_{x_i}(\sigma_1\psi_i)^j}\varphi_i^j$$

In the middle step the scalar product is meant on $L^2(\mathbb{R}^+; \mathbb{C}^2)$. At this point we use the fact that integration *per partes* can be used on Sobolev spaces H^1. We obtain

$$\sum_{i=1}^{n} \left(\langle \psi_i | (-\mathrm{i}d_{x_i} \otimes \sigma_1)\, \varphi_i\rangle - \mathrm{i}\left(\overline{\psi_i^2(0)}\varphi_i^1(0) + \overline{\psi_i^1(0)}\varphi_i^2(0) \right) \right).$$

If we sum each term over i separately, we get the desired result.

$$\begin{aligned} \sum_{i=1}^{n} \langle \psi_i | (-\mathrm{i}d_{x_i} \otimes \sigma_1)\, \varphi_i\rangle &= \langle \Psi | D_{max}\Phi\rangle \\ \sum_{i=1}^{n} -\mathrm{i}\overline{\psi_i^2(0)}\varphi_i^1(0) &= \langle \Gamma_2\Psi|\Gamma_1\Phi\rangle \\ \sum_{i=1}^{n} -\mathrm{i}\overline{\psi_i^1(0)}\varphi_i^2(0) &= -\langle \Gamma_1\Psi|\Gamma_2\Phi\rangle \end{aligned}$$

The second condition to verify is that the map $(\Gamma_1, \Gamma_2) : \mathrm{Dom}\, D_{max} \to \mathscr{G} \oplus \mathscr{G}$ is surjective. From

$$(\Gamma_1, \Gamma_2)\Psi = \left(\begin{pmatrix} \psi_1^1(0) & \cdots & \psi_n^1(0) \end{pmatrix}^T, \begin{pmatrix} \mathrm{i}\psi_1^2(0) & \cdots & \mathrm{i}\psi_n^2(0) \end{pmatrix}^T \right)$$

we see that if we take any function from $H^1(\mathbb{R}^+; \mathbb{C})$ with non-zero trace, e.g. take $\tilde{\psi}(x) = \mathrm{e}^{-x}$, and define $\Psi_j^i \in \mathrm{Dom}\, D_{max}$ with all entries zero except on the edge e_i the j-th spinor component equal to $\tilde{\psi}$, then images $(\Gamma_1, \Gamma_2)\Psi_j^i$ form a basis in $\mathscr{G} \oplus \mathscr{G}$, hence the map (Γ_1, Γ_2) is surjective.

Last condition to check is a density of $\mathrm{Ker}\,(\Gamma_1, \Gamma_2)$ in $\mathscr{H}$. This follows immediately from the observation that the test functions (smooth compactly supported) are dense in L^2 space and

$$\bigoplus_{i=1}^{n} \mathscr{D}(\mathbb{R}^+; \mathbb{C}^2) \subset \bigoplus_{i=1}^{n} H_0^1(\mathbb{R}^+; \mathbb{C}^2) = \mathrm{Ker}\,(\Gamma_1, \Gamma_2) \subset \bigoplus_{i=1}^{n} L^2(\mathbb{R}^+; \mathbb{C}^2) = \mathscr{H}.$$

This concludes the proof that $(\mathscr{G}, \Gamma_1, \Gamma_2)$ is a boundary triple for D_{max}. □

The problem of self-adjoint extensions is now fully addressed by [1, Theorem 1.12], in our case namely by the following corollary.

Corollary 2. There is a one-to-one correspondence between all self-adjoint relations Λ in $\mathscr{G} = \mathbb{C}^n$ and all self-adjoint extensions of D_0 given by $\Lambda \leftrightarrow D_\Lambda$, where D_Λ is the restriction of D_{max} to the subset

$$\mathrm{Dom}\, D_\Lambda = \{\Psi \in \mathrm{Dom}\, D_{max} | (\Gamma_1\Psi, \Gamma_2\Psi) \in \Lambda\}.$$

Remark 3. In the following section, one significant self-adjoint extension will become helpful. Namely, a one defined as the restriction of D_{max} to $\mathrm{Ker}\,\Gamma_1$, or equivalently D_Λ where $\Lambda = 0 \oplus \mathscr{G}$. To simplify the notation, we continue to write $D^{0,\mathscr{G}}$ for $D_{0\oplus\mathscr{G}}$.

4 Krein's resolvent formula

The Krein resolvent formula relates a spectrum and a resolvent of any self-adjoint extension to the spectrum and the resolvent of one particular extension which is chosen to be the operator $D^{0,\mathscr{G}}$ from the remark 3. Besides the boundary triple, it is needed to introduce so called Krein Γ-field and $\mathscr{Q}$-function. This will be the content of the following paragraphs.

Firstly, denote the defect subspaces of D_0 as $N_z = \mathrm{Ker}\,(D_{max} - z)$. Consequently, the subspace is the solution of $(D_{max} - z)\Psi = 0$ in the domain of D_{max}. Using notation as in (2) and putting $\psi_i = \begin{pmatrix} \psi_i^1 & \psi_i^2 \end{pmatrix}^T$, the equation can be written as the following.

$$\begin{aligned} &\forall i \in \{1, \ldots, n\} \\ &(-\mathrm{i}d_{x_i} \otimes \sigma_1 - z)\,\psi_i = 0 \qquad /\mathrm{i}\sigma_1 \cdot \\ &d_{x_i}\psi_i = \mathrm{i}\sigma_1 z \psi_i \end{aligned}$$

The solution is

$$\psi_i(x) = \exp(\mathrm{i}\sigma_1 z x)\begin{pmatrix}\omega_i^1\\ \omega_i^2\end{pmatrix} = \left[\mathrm{e}^{\mathrm{i}zx}\frac{1}{2}(I+\sigma_1) + \mathrm{e}^{-\mathrm{i}zx}\frac{1}{2}(I-\sigma_1)\right]\begin{pmatrix}\omega_i^1\\ \omega_i^2\end{pmatrix}$$

where ω_i^j are complex integration constants. To be in $\operatorname{Dom} D_{max}$, the solution has to be square integrable. Therefore, for $\Im z > 0$ the term with $\mathrm{e}^{-\mathrm{i}zx}$ has to be missing and hence $\forall i : \omega_i^1 = \omega_i^2$, and analogously for $\Im z < 0$ the term with $\mathrm{e}^{\mathrm{i}zx}$ has to be missing and hence $\forall i : \omega_i^1 = -\omega_i^2$. Finally, for a real z only the zero solution fulfils the requirements. To summarise, assuming $\Im z \neq 0$ the defect subspace N_z is the following set (involving the G matrix defined in (4)).

$$N_z = \{\mathrm{e}^{\mathrm{i}\operatorname{sgn}(\Im z)zx} G^T_{1,\operatorname{sgn}(\Im z)}\alpha \mid \alpha \in \mathbb{C}^n\} \tag{5}$$

Notice now, how Γ_1 maps elements from N_z to $\mathbb{C}^n$. The restriction is injective, therefore the inverse map can be denoted by $\gamma(z) = (\Gamma_1|_{N_z})^{-1} : \mathscr{G} \to N_z$ and acting as

$$\forall z \in \mathbb{C}\setminus\mathbb{R}, \forall \alpha \in \mathscr{G} = \mathbb{C}^n : \quad \gamma(z)\alpha = x \mapsto \mathrm{e}^{\mathrm{i}\operatorname{sgn}(\Im z)zx} G^T_{1,\operatorname{sgn}(\Im z)}\alpha. \tag{6}$$

The map γ is called the Krein Γ-field.

Finally, define the $\mathscr{Q}$-function as a composition of γ and Γ_2

$$\forall z \in \mathbb{C}\setminus\mathbb{R} : Q(z) = \Gamma_2|_{N_z} \circ \gamma(z) : \mathscr{G} \to \mathscr{G}.$$

Specifically

$$\forall z \in \mathbb{C}\setminus\mathbb{R}, \forall \alpha \in \mathscr{G} = \mathbb{C}^n : \quad Q(z)\alpha = \mathrm{i}\operatorname{sgn}(\Im z)\alpha.$$

The Krein resolvent formula, which was the main objective of this paper, is in our case according to [1, Theorem 1.29] the following corollary.

Corollary 4 (Krein resolvent formula)**.** For any self-adjoint linear relation Λ in $\mathscr{G}$, it holds

1. $\forall z \in \rho(D^{0,\mathscr{G}}) : \quad \operatorname{Ker}(D_\Lambda - z) = \gamma(z)\operatorname{Ker}(Q(z) - \Lambda)$,
2. $\forall z \in \rho(D^{0,\mathscr{G}}) \cap \rho(D_\Lambda) : \quad 0 \in \rho(Q(z) - \Lambda)$ and
   $(D^{0,\mathscr{G}} - z)^{-1} - (D_\Lambda - z)^{-1} = \gamma(z)\,(Q(z) - \Lambda)^{-1}\,\gamma^*(\overline{z})$,
3. $\sigma(D_\Lambda) \setminus \sigma(D^{0,\mathscr{G}}) = \{z \in \rho(D^{0,\mathscr{G}}) \mid 0 \in \sigma\,(Q(z) - \Lambda)\}$.

5 Isomorphism to operator acting on real line

The interest of this section is a special case of star-shaped graph with exactly two edges. This graph resembles the real line, the first edge associated with negative real numbers and the second edge associated with positive real numbers. Consequently, the vertex should be mapped to the zero of the real line. The idea is to compare the results for star-shaped graph with the point interaction on the real line [3].

From the point of view of metric spaces, it is clear how the isomorphism between the graph and the line should be defined. However, an isomorphism of Hilbert spaces that

also respects studied Dirac operator is not as trivial as it could seem. This is also the difference in comparison to the Schrödinger operator (minus Laplacian) because in that case the naive isomorphism works.

To state the problem rigorously, put $\mathscr{H}_{graph} = L^2(\mathbb{R}^+;\mathbb{C}^2) \oplus L^2(\mathbb{R}^+;\mathbb{C}^2)$ as the special case of the general Hilbert space (1). Also put $\mathscr{H}_{line} = L^2(\mathbb{R}\setminus\{0\};\mathbb{C}^2)$ as Hilbert space on the real line. The exclusion of the zero will become clear as soon as domains of Dirac operators will be defined.

According to (3), the maximal Dirac operator in the space $\mathscr{H}_{graph}$ is given by $D_{graph} = (-\mathrm{i}d_{x_L} \otimes \sigma_1) \oplus (-\mathrm{i}d_{x_R} \otimes \sigma_1)$ where the first edge is called the *left* with a variable x_L and the second edge is called the *right* with a variable x_R. The domain of the D_{graph} is $\operatorname{Dom} D_{graph} = H^1(\mathbb{R}^+;\mathbb{C}^2) \oplus H^1(\mathbb{R}^+;\mathbb{C}^2)$.

The maximal Dirac operator on the real line is defined analogously $D_{line} = -\mathrm{i}d_x \otimes \sigma_1$ with the domain $\operatorname{Dom} D_{line} = H^1(\mathbb{R}\setminus\{0\};\mathbb{C}^2)$. This is consistent with the point interaction procedure in [3].

Now, the objective is to find a bijection $J : \mathscr{H}_{graph} \to \mathscr{H}_{line}$ such that the following diagram commutes in the sense that

$$\forall \Psi \in \operatorname{Dom} D_{graph} : \qquad (D_{line} \circ J)(\Psi) = (J \circ D_{graph})(\Psi).$$

$$\begin{array}{ccc} \operatorname{Dom} D_{graph} & \xrightarrow{D_{graph}} & \mathscr{H}_{graph} \\ {\scriptstyle J}\downarrow & & \downarrow{\scriptstyle J} \\ \operatorname{Dom} D_{line} & \xrightarrow{D_{line}} & \mathscr{H}_{line} \end{array}$$

The claim is that a possible J is defined as the following.

$$\begin{gathered} \forall \psi_L, \psi_R \in L^2(\mathbb{R}^+;\mathbb{C}^2), \text{ i.e. } \psi_L \oplus \psi_R \in \mathscr{H}_{graph} \\ J(\psi_L \oplus \psi_R) = x \in \mathbb{R}\setminus\{0\} \mapsto \begin{cases} \psi_R(x) & \text{for } x > 0 \\ \overline{\psi_L(-x)} & \text{for } x < 0 \end{cases} \end{gathered} \tag{7}$$

The important thing to notice is the complex conjugation of ψ_L. The inverse map is given by the following.

$$\begin{aligned} \forall \varphi \in L^2(\mathbb{R}\setminus\{0\};\mathbb{C}^2) \qquad & J^{-1}\varphi = \varphi_L \oplus \varphi_R \\ \text{where } \forall x \in \mathbb{R}^+ : \qquad & \varphi_L(x) = \overline{\varphi(-x)} \\ & \varphi_R(x) = \varphi(x) \end{aligned}$$

It remains to compute both ways of composing J and the appropriate Dirac operator. Let $\psi_L \oplus \psi_R$ be any element from $\operatorname{Dom} D_{graph}$. Then the application of the graph Dirac operator and subsequently J gives

$$(J \circ D_{graph})(\psi_L \oplus \psi_R) = J(-\mathrm{i}\sigma_1\psi_L' \oplus -\mathrm{i}\sigma_1\psi_R') = x \mapsto \begin{cases} -\mathrm{i}\sigma_1\psi_R'(x) & \text{for } x > 0 \\ +\mathrm{i}\sigma_1\overline{\psi_L'(-x)} & \text{for } x < 0 \end{cases}.$$

The other way with the line Dirac operator yields

$$(D_{line}\circ J)(\psi_L\oplus\psi_R) = D_{line}\left(x \mapsto \begin{cases} \psi_R(x) & \text{for } x > 0 \\ \overline{\psi_L(-x)} & \text{for } x < 0 \end{cases}\right) = x \mapsto \begin{cases} -\mathrm{i}\sigma_1\psi_R'(x) & \text{for } x > 0 \\ +\mathrm{i}\sigma_1\overline{\psi_L'(-x)} & \text{for } x < 0 \end{cases}.$$

Hence

$$J \circ D_{graph} = D_{line} \circ J \qquad \text{on } \operatorname{Dom} D_{graph}.$$

It could be said that the map J preserves the Dirac operator while the configuration space is described differently. However, the map does not preserve the structure of vector space as it is not linear nor anti-linear. The same map also preserves the Schrödinger operator (minus Laplacian).

However, the operators D_{graph} and D_{line} are not self-adjoint. The self-adjoint realisations are their restrictions described by certain boundary conditions. The operator D_{graph} has the boundary condition from the corollary 2. According to [1, Theorem 1.2] every self-adjoint relation Λ is uniquely parametrised by an unitary operator U on $\mathscr{G} = \mathbb{C}^2$ such that

$$\Lambda = \{(w_1, w_2) \in \mathbb{C}^2 \oplus \mathbb{C}^2 | \mathrm{i}(I_2 + U)w_1 = (I_2 - U)w_2\}. \tag{8}$$

On the other hand, one type of delta interaction is in [3, Equation 1.15] described by the equations for one-sided limits at the zero.

$$\begin{aligned} \chi \in H^1(\mathbb{R} \setminus \{0\}; \mathbb{C}^2) \qquad & \theta_1, \theta_2 \in [0, 2\pi) \\ \chi^1(0_+)(1 - \mathrm{e}^{\mathrm{i}\theta_1}) &= \chi^2(0_+)(1 + \mathrm{e}^{\mathrm{i}\theta_1}) \\ \chi^1(0_-)(1 + \mathrm{e}^{\mathrm{i}\theta_2}) &= \chi^2(0_-)(-1 + \mathrm{e}^{\mathrm{i}\theta_2}). \end{aligned} \tag{9}$$

The rest of the section demonstrates how every point interaction of the form of (9) is described by self-adjoint restriction of D_{graph} on (8) at the vertex. The claim is that the right parametrization of Λ in (8) is

$$U = \begin{pmatrix} \mathrm{e}^{\mathrm{i}\theta_2} & 0 \\ 0 & -\mathrm{e}^{\mathrm{i}\theta_1} \end{pmatrix}. \tag{10}$$

Plugging U into (8) with $w_1 = \Gamma_1\Psi, w_2 = \Gamma_2\Psi$ with $\Psi \in \operatorname{Dom} D_{graph}$ yields

$$\begin{pmatrix} 1 + \mathrm{e}^{\mathrm{i}\theta_2} & 0 \\ 0 & 1 - \mathrm{e}^{\mathrm{i}\theta_1} \end{pmatrix} \begin{pmatrix} \psi_L^1(0) \\ \psi_R^1(0) \end{pmatrix} = \begin{pmatrix} 1 - \mathrm{e}^{\mathrm{i}\theta_2} & 0 \\ 0 & 1 + \mathrm{e}^{\mathrm{i}\theta_1} \end{pmatrix} \begin{pmatrix} \psi_L^2(0) \\ \psi_R^2(0) \end{pmatrix}. \tag{11}$$

Using the bijection J (7)

$$\forall j \in \{1, 2\}: \qquad \psi_L^j(0) = \overline{\psi^j(0_-)} \text{ and } \psi_R^j(0) = \psi^j(0_+)$$

the second row of (11) becomes exactly the first condition (9). The first row is

$$(1 + \mathrm{e}^{\mathrm{i}\theta_2})\overline{\psi^1(0_-)} = (1 - \mathrm{e}^{\mathrm{i}\theta_2})\overline{\psi^2(0_-)}.$$

To get the form of (9) complex conjugation of the equation needs to be multiplied by $(1 + \mathrm{e}^{\mathrm{i}\theta_2})/(1 + \mathrm{e}^{-\mathrm{i}\theta_2})$. If the numerator or denominator was zero then it is not necessary (or even possible) because one of the sides was already zero and in the correct form. This correspondence concludes the section.

6 Conclusion

After introducing the star-shaped graph and assigning to it the Hilbert space $\mathscr{H}$ of L^2 functions, it was proceeded by studying the Dirac operators. To find all self-adjoint realisations, the theory of boundary triples was used. It turned out that every self-adjoint Dirac operator corresponds to a self-adjoint linear relation in the auxiliary Hilbert space $\mathscr{G} = \mathbb{C}^n$. In the next section, the Krein resolvent formula was stated for the studied problem.

Finally, the bijection was provided between star-shaped relativistic quantum graph with exactly two edges and relativistic point interaction in one dimension.

The aim of the next research should be generalizing the results for more general underlying graphs. The other suggestion is to release the condition of self-adjointness and study operators which adjoint is only similar to them – quasi-self-adjointness. The motivation is to obtain analogue results as for the Laplace operator in [2].

References

[1] J. Brüning, V. Geyler, and K. Pankrashkin. *Spectra of self-adjoint extensions and applications to solvable schrödinger operators.* Reviews in Mathematical Physics **20** (2008), 1–70.

[2] A. Hussein, D. Krejčiřík, and P. Siegl. *Non-self-adjoint graphs.* Transactions of the American Mathematical Society **367** (2014), 2921–2957.

[3] V. Růžek. *One-dimensional relativistic point interactions – approximation by regular potentials, application to models of the dirac materials.* Master's thesis, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, (2021).

Dynamic Decision-Making and Preferences Quantification with Meta Closed-Loop

Tereza Siváková
sivakter@cvut.cz

study programme: Mathematical Engineering
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. This contribution addresses the challenging task of specifying user preferences in decision-making problems [2]. The average user is not educated in decision-making processes, which can lead to an incomplete expression of their preferences or unrealistic and contradictory preferences. This study uses a fully probabilistic design [3] (FPD) framework, which effectively models the closed-loop between the user and a system in decision-making processes. Within the FPD framework, an ideal probability density is introduced, assigning high probabilities to preferred behaviors and low probabilities to undesirable ones. The optimal decision policy is then determined by minimizing the Kullback-Leibler divergence between the real and ideal probability densities.

As previously mentioned, preferences can become particularly complex, especially when users have preferences for both states and actions. There is a parameter of weight that has to be estimated to balance these two preferences. Estimating this parameter is a very challenging task because users may not be able to express how much they prefer one option over another, and predicting the weight without knowledge of the system is difficult. This challenge is resolved by adding another (meta) closed-loop. The user observes sequences of states and actions and then rates how much they like them using a grading system similar to that used in schools. In doing so, the user can remotely adjust the parameter values. Based on this feedback, the parameters of the main closed-loop are fine-tuned.

To validate the effectiveness of the meta closed-loop theory, we have developed a Python-based web application. This application empowers users to actively influence the results of the closed-loop and experiment with the theory, allowing us to further test and refine our approach. http://nebula.utia.cas.cz/

Keywords: Adaptive, agent, Bayes' rule, Decision making, Preference elicitation,

Abstrakt. Tento příspěvek se zabývá náročným úkolem *specifikovat uživatelovy preference v rozhodovacích problémech [2].* Běžný uživatel není v matematice rozhodovacích procesů vzdělaný a to může to vést k neúplnému vyjádření preferencí nebo nereálným a protichůdným preferencím. Nástroj použitý v tomto příspěvku se nazývá plně pravděpodobnostního návrh [3] (PPN), který efektivně modeluje uzavřenou smyčku mezi uživatelem a systémem v rozhodovacích procesech. V rámci PPN je zavedena ideální hustota pravděpodobnosti, která přiřazuje vysoké pravděpodobnosti preferovanému chování a nízké pravděpodobnosti nežádoucímu. Optimální rozhodovací politika je pak určena minimalizací Kullback-Leiblerovy divergence mezi skutečnou a ideální hustotou pravděpodobnosti.

Jak bylo uvedeno výše, preference mohou být protichůdné, zejména pokud uživatel může

preferovat jak stavy tak i akce. Existuje parametr váhy, který vyvažuje tyto dvě preference a je třeba ho odhadnout. Odhadnout tento parametr je ale velmi obtížné, jelikož uživatel nedokáže vyjádřit, jak moc preferuje jedno před druhým. Také je obtížné předpovědět váhu bez znalosti systému. Tento příspěvek vyřešil tento problém přidáním další uzavřené smyčky. Uživatel sleduje posloupnosti stavů a akcí a následně hodnotí, jak se mu líbí, pomocí školních známek. Tím uživatel vlastně nepřímo mění hodnoty parametrů, jelikož na základě zpětné vazby se ladí parametry hlavní uzavřené smyčky.

Pro testování teorie meta uzavřené smyčky byla vyvinuta webová aplikace naprogramovaná v Pythonu, která uživatelům nabízí možnost aktivně ovlivnit výsledky uzavřené smyčky a tuto teorii tak vyzkoušet. http://nebula.utia.cas.cz/

Klíčová slova: Adaptivní, agent, Bayesovské pravidlo, rozhodování, preference

Full paper: M. Kárný and T. Siváková, *Model-Based Preference Quantification*, Automatica vol.156, 111185, 2023. DOI: 10.1016/j.automatica.2023.111185.

References

[1] T. Siváková and M. Kárný, *Experiments with the User's Feedback in Preference Elicitation*, AIxIA – 21st Int. Conf. of the Italian Association for AI, 2022, Vol-3463 (Udine, IT, 20221127)

[2] M.L.Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, 2005.

[3] M.Kárný, *Towards fully probabilistic control design* , Automatica vol.32, 12, 1996, p. 1719-1722

Quantum Walk State Transfer on a Hypercube*

Stanislav Skoupý
Stanislav.Skoupy@fjfi.cvut.cz

study programme:
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Martin Štefaňák, Department of Physics, Faculty of Nuclear Science and Physical Engineering, CTU in Prague

Abstract. We investigate state transfer on a hypercube by means of a discrete-time quantum walk where the sender and the receiver vertices are marked by a loops with optimally chosen weight. First, we analyze search for a single marked vertex, which can be used for state transfer between arbitrary vertices by switching the weighted loop from the sender to the receiver after one run-time. Next, state transfer between antipodal vertices is considered. We show that one can tune the weight of the loop to achieve state transfer with high fidelity in shorter run-time in comparison to the state transfer with a switch. Finally, we investigate state transfer between vertices of arbitrary distance. It is shown that when the distance between the sender and the receiver is at least 2, the results derived for the antipodes are well applicable. If the sender and the receiver are direct neighbours the evolution follows a slightly different course. Nevertheless, state transfer with high fidelity is achieved in the same run-time.

Keywords: quantum walk, search algorithm, state transfer algorithm, hypercube

Abstrakt. Zkoumáme přenos stavu na hypekrychly dosaženého pomocí kvantové procházky v diskrétním čase, kde na vrcholech odesílatele a příjemce je umístěna smyčka s optimálně zvolenou váhou. Nejprve analyzujeme vyhledávání jednoho označeného vrcholu, které může být použito pro přenos stavu mezi libovolnými vrcholy, když přepneme váženou smyčku z odesílatele na příjemce po jednom běhu vyhledávání. Dále uvažujeme přenos stavu mezi protějšími vrcholy. Ukážeme, že lze zvolit váhu smyčky, tak aby byl dosažen přenos stavu s vysokou pravděpodobností přenosu v kratším čase v porovnání s přenosem stavu s přepnutím. Nakonec zkoumáme přenos stavu mezi vrcholy libovolné vzdálenosti. Ukážeme, že když vzdálenost mezi odesílatelem a příjemcem je alespoň 2, můžeme aplikovat výsledky odvozené pro protilehlé vrcholy. Pokud odesílatel a příjemce jsou sousedé, evoluce probíhá trochu jinak, ale přenos stavu s vysokou pravděpodobností přenosu je dosažen ve stejném čase.

Klíčová slova: kvantová procházka, vyhledávací algoritmus, algoritmus na přesnos stavu, hyperkrychle

Plná verze: Quantum walk state transfer on a hypercube, Martin Štefaňák and Stanislav Skoupý, (2023), Phys. Scr. 98 104003

*This work was supported from Student Grant Competition of Czech Technical University in Prague under Grant SGS22/181/OHK4/3T/14 and project CAAS.

New Families of Orthogonal Polynomials Generated from the Level One Solutions of the Heun Equation*

Patrik Šnauko
snaukpat@fjfi.cvut.cz

study programme: Mathematical Engineering
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: František Štampach, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. In the following text we would like to introduce motivation and first results of investigating new families of orthogonal polynomials which are connected to the solutions of the Heun equation. We are especially interested in the so-called *level N solutions.* Devices which are frequently used in the following text are *the Jacobi elliptic functions* $(\mathrm{sn}(z), \mathrm{cn}(z), \mathrm{dn}(z))$ and the four *theta functions* $(H(z), H_1(z), \Theta(z), \Theta_1(z))$.

Keywords: Heun equation, Jacobi matrix and operator, orthogonal polynomials

Abstrakt. V tomto příspěvku bychom rádi představili motivaci a několik prvních výsledků nalézání nových rodin ortogonálních polynomů, které jsou spojeny s řešením Heunovy diferenciální rovnice. Speciálně nás zajímají tzv. *řešení N-té úrovně.* Často používaným nástrojem jsou tzv. *Jacobiho eliptické funkce* $(\mathrm{sn}(z), \mathrm{cn}(z), \mathrm{dn}(z))$ a čtyři *theta funkce* $(H(z), H_1(z), \Theta(z), \Theta_1(z))$.

Klíčová slova: Heunova diferenciální rovnice, Jacobiho matice a operátor, ortogonální polynomy

1 Motivation and preliminaries

In 1960, Leonard Carlitz introduced six new families of orthogonal polynomials in [4]. Let us ilustrate our motivation on the particular example of the family $\{f_n(x)\}_{n=0}^{\infty}$ (we assume the Carlitz notation). These polynomials are given by the three-terms recurrence

$$\begin{aligned} f_{n+1}(x) &= \left(x + (k^2+1)(2n+1)^2\right) f_n(x) - k^2(2n-1)(2n)^2(2n+1) f_{n+1}(x), \\ f_0(x) &= 1, \quad f_1(x) = x + k^2 + 1, \end{aligned} \tag{1}$$

with $k \in (0,1)$. Carlitz has shown that the corresponding measure of orthogonality is

$$\mu_f = \sum_{n=1}^{\infty} w_n^{(f)} \delta_{x_n^{(f)}},$$

with

$$x_n^{(f)} = \frac{(2n+1)^2\pi^2}{16K^2} - \frac{1}{4} + \frac{5}{4}k^2, \quad w_n^{(f)} = \frac{(2n+1)\pi^2}{kK^2} \frac{q^{n+\frac{1}{2}}}{1-q^{2n+1}}. \tag{2}$$

*This work has been kindly supported by the GAČR EXPRO grant No. 20-17749X.

Numbers K and q are commonly used constants in the theory of the Jacobi elliptic functions, see [7]. L. Carlitz derived the measure thanks to the good knowledge of these functions. In my doctoral thesis we desire to derive and especially extend results from [4] in, daresay, more general way.

To introduce the connection between our topic and the family $\{f_n(x)\}_{n=0}^{\infty}$, we need to summarize some facts about the Heun equation, orthogonal polynomials and the link between them.

1.1 The Heun equation and the level N solutions

The second order differential equation

$$F''(w) + \left(\frac{\gamma}{w} - \frac{\delta}{1-w} - \frac{\epsilon k^2}{1-k^2 w}\right) F'(w) + \frac{s + \alpha\beta k^2 w}{w(1-w)(1-k^2 w)} F(w) = 0, \tag{3}$$

where $k \in (0,1)$ with additional condition

$$\alpha + \beta + 1 = \gamma + \delta + \epsilon. \tag{4}$$

is called *the Heun equation*. Solution F in the neighbourhood of 0 with $F(0) = 0$ and $F'(0) = 1$ exists and is called *the Heun function*. The Heun function will be denoted by

$$\mathrm{Hn}(k^2, s; \alpha, \beta, \gamma, \delta; w).$$

Note the useful transformation of equation (3) provided by substitution for $w = \mathrm{sn}^2(z)$ and setting $v(z) = F(w)$

$$v''(z) + \left((2\gamma-1)\frac{\mathrm{cn}(z)\mathrm{dn}(z)}{\mathrm{sn}(z)} - (2\delta-1)\frac{\mathrm{sn}(z)\mathrm{dn}(z)}{\mathrm{cn}(z)} - (2\epsilon-1)k^2\frac{\mathrm{sn}(z)\mathrm{cn}(z)}{\mathrm{dn}(z)}\right) v'(z) \tag{5}$$
$$+4(s + \alpha\beta k^2 \mathrm{sn}^2(z))v(z) = 0.$$

Meromorphic solutions of (5) are of interest. Valent states in [10] that the necessary conditions for meromorphy are

$$\gamma = \frac{1}{2} - m_1, \;\; \delta = \frac{1}{2} - m_2, \;\; \epsilon = \frac{1}{2} - m_3, \;\; M := m_1 + m_2 + m_3$$

and

$$\alpha = -\frac{1}{2}(m_0 + M), \;\; \beta = \frac{1}{2}(m_0 - M + 1), \;\; N := m_0 + M$$

for $(m_0, m_1, m_2, m_3) \in \mathbb{Z}^4$. Valent also refers in [10] that these conditions are sufficient for the meromorphy as well. Equation (5) then takes the form

$$v''(z) + 2\left(-m_1\frac{\mathrm{cn}(z)\mathrm{dn}(z)}{\mathrm{sn}(z)} + m_2\frac{\mathrm{sn}(z)\mathrm{dn}(z)}{\mathrm{cn}(z)} + m_3 k^2\frac{\mathrm{sn}(z)\mathrm{cn}(z)}{\mathrm{dn}(z)}\right) v'(z) \tag{6}$$
$$+(4s + N(N - 2m_0 - 1)k^2\mathrm{sn}^2(z))v(z) = 0.$$

In the case that $(m_0, m_1, m_2, m_3) \in \mathbb{N}_0^4$, we call the respective solutions *level N-solutions* of the Heun equation (6).

1.2 The Jacobi matrix and orthogonal polynomials

The Jacobi matrix is a semi-infinite tridiagonal matrix of the form

$$\mathcal{J} = \begin{pmatrix} \beta_0 & \alpha_0 & 0 & 0 & 0 & \dots \\ \alpha_0 & \beta_1 & \alpha_1 & 0 & 0 & \dots \\ 0 & \alpha_1 & \beta_2 & \alpha_2 & 0 & \dots \\ 0 & 0 & \alpha_2 & \beta_3 & \alpha_3 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{7}$$

$\mathbb{C}^\infty$ denotes a linear space of semi-infinite column vectors with components indexed by $\mathbb{N}_0$ and $\mathbb{C}^{\infty,\infty}$ denotes the linear space of semi-infinite matrices with entries indexed by $\mathbb{N}_0$. Hilbert space ℓ^2 may be seen as the vector subspace of $\mathbb{C}^\infty$.

Every band matrix (in particular, every tridiagonal matrix) $\mathcal{A} \in \mathbb{C}^{\infty,\infty}$ becomes naturally a linear operator on $\mathbb{C}^\infty$.

Denote $\mathbf{P}(x) := (1, P_1(x), \dots, P_n(x), \dots)^T$. Then the formal eigenvector equation

$$\mathcal{J}\mathbf{P}(x) = x\mathbf{P}(x)$$

defines an orthogonal polynomials sequence (OPS) $\{P_n(x)\}$. The above equation can be rewritten as

$$\begin{aligned} P_0(x) &= 1, \\ \alpha_0 P_1(x) + (\beta_0 - x)P_0(x) &= 0, \\ \alpha_n P_{n+1}(x) + (\beta_n - x)P_n(x) + \alpha_{n-1}P_{n-1}(x) &= 0, \quad \text{for } n \geq 1. \end{aligned} \tag{8}$$

Matrix $\mathcal{J}$ can be treated as an operator $\dot{J}$ on ℓ^2 defined by

$$\mathrm{Dom}(\dot{J}) = \mathrm{span}\{\mathbf{e}_n\}_{n=0}^\infty, \qquad \dot{J}\mathbf{f} = \mathcal{J}\mathbf{f}, \quad \text{for } \mathbf{f} \in \mathrm{Dom}(\dot{J}). \tag{9}$$

Clearly, operator $\dot{J}$ is symmetric. Thus it has at least one self-adjoint extension. One of them can be defined as

$$\mathrm{Dom}(J) = \left\{\mathbf{f} \in \ell^2; \mathcal{J}\mathbf{f} \in \ell^2\right\}, \qquad J\mathbf{f} = \mathcal{J}\mathbf{f}, \quad \text{for } \mathbf{f} \in \mathrm{Dom}(J). \tag{10}$$

Operator J is maximal in sense of inclusion of the domain. Moreover

$$J = \dot{J}^*.$$

If it happens that the moment problem for polynomials $\{P_n(x)\}_{n=0}^\infty$ is determinate (see [1]), J is the only self-adjoint extension of $\dot{J}$ and hence $\dot{J}$ is essentially self-adjoint.

In the case that $\alpha_n > 0$, we can define polynomials $\{p_n(x)\}_{n=0}^\infty$ by

$$p_n(x) = \left(\prod_{k=0}^{n-1} \alpha_k\right) P_n(x). \tag{11}$$

Polynomials $p_n(x)$ are orthogonal, monic and they satisfy the three-terms recurrence

$$p_{n+1}(x) = (x - \beta_n)p_n(x) - \alpha_{n-1}^2 p_{n-1}(x). \tag{12}$$

1.3 The Heun function vs. orthogonal polynomials

The following results come from [10]. Set coefficients of $\mathcal{J}$

$$\alpha_n := \sqrt{\lambda_n \nu_{n+1}}, \qquad \beta_n := \lambda_n + \nu_n + \gamma_n \tag{13}$$

with

$$\lambda_n = k^2(n+\alpha)(n+\beta), \tag{14}$$
$$\nu_n = n(n+\gamma-1), \tag{15}$$
$$\gamma_n = (1-k^2)\delta n. \tag{16}$$

Numbers α, β, γ and δ are coefficients of the Heun equation (3). Next, define

$$F(w) := \sum_{n=0}^{\infty} (-1)^n \sqrt{\frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\nu_1 \nu_2 \dots \nu_n}} P_n(s + \alpha\beta k^2) w^n. \tag{17}$$

Then

$$F(w) = \mathrm{Hn}(k^2, s; \alpha, \beta, \gamma, \delta; w).$$

Moreover, with these settings, the corresponding moment problem is determinate, thus operator $\dot{J}$ defined by this matrix is essa. Using polynomials $p_n(x)$ instead of $P_n(x)$ in (17) yields

$$\mathrm{Hn}(k^2, s; \alpha, \beta, \gamma, \delta; w) = \sum_{n=0}^{\infty} \frac{(-1)^n}{\nu_1 \nu_2 \dots \nu_n} p_n(s + \alpha\beta k^2) w^n. \tag{18}$$

1.4 The first comeback to the motivation

Let us focus on the special case of equation (6) with $m_0 = m_1 = m_2 = m_3 = 0$. It corresponds to the Heun coefficients $(\alpha, \beta, \gamma, \delta, \epsilon) = \left(0, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$ and reads

$$v''(z) + 4sv(z) = 0.$$

By solving this trivial equation one gets

$$\mathrm{Hn}\left(k^2, s; 0, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}; w\right) = \cos\left(2\sqrt{s}\ \mathrm{sn}^{-1}(\sqrt{w})\right). \tag{19}$$

It is readily seen that (19) is the level 0 solution of the Heun equation.

On the other hand, comparing coefficients of the three-terms recurrence in (1) with those in (12) yields that

$$p_n(x + \alpha\beta k^2) = \frac{(-1)^n}{4^n} f_n(-4x - 1 - k^2), \quad n \in \mathbb{N}_0 \tag{20}$$

for three admissible choices of parameters

$$(\alpha, \beta, \gamma, \delta, \epsilon) \in \left\{ \left(\frac{1}{2}, 1, \frac{3}{2}, \frac{1}{2}, \frac{1}{2}\right), \left(1, \frac{3}{2}, \frac{1}{2}, \frac{3}{2}, \frac{3}{2}\right), \left(\frac{1}{2}, \frac{3}{2}, 1, 1, 1\right) \right\}.$$

None of these corresponds to the level 0 solution of the Heun equation which was derived above. However, according to [9] the following identities hold true.

$$\frac{\mathrm{d}}{\mathrm{d}w}\mathrm{Hn}\left(k^2,s;0,\beta,\gamma,\delta;w\right)=-\frac{s}{\gamma}\mathrm{Hn}\left(k^2,s-\gamma-\delta-k^2(\gamma+\epsilon);2,\beta+1,\gamma+1,\delta+1;w\right)$$

$$\begin{aligned}\mathrm{Hn}(k^2,s;\alpha,\beta,\gamma,\delta;w)&=(1-w)^{1-\delta}(1-k^2w)^{1-\epsilon}\times\\&\times\mathrm{Hn}\left(k^2,s+\gamma(\delta-1+k^2(\epsilon-1));-\alpha+\gamma+1,-\beta+\gamma+1,\gamma,2-\delta;w\right).\end{aligned}$$

Applying these two identities to the level 0 solution yields

$$\mathrm{Hn}\left(k^2,s-\frac{1}{4}-\frac{k^2}{4};\frac{1}{2},1,\frac{3}{2},\frac{1}{2};w\right)=\frac{\sin\left(2\sqrt{s}\ \mathrm{sn}^{-1}(\sqrt{w})\right)}{2\sqrt{s}\sqrt{w}}.$$

In a view of (18) and (20) one has

$$\mathrm{Hn}\left(k^2,s-\frac{1}{4}-\frac{k^2}{4};\frac{1}{2},1,\frac{3}{2},\frac{1}{2};w\right)=\sum_{n=0}^{\infty}\frac{f(-4s)}{(2n+1)!}w^n.$$

Thus,

$$\sum_{n=0}^{\infty}\frac{f(-x)}{(2n+1)!}w^n=\frac{\sin\left(\sqrt{x}\ \mathrm{sn}^{-1}(\sqrt{w})\right)}{\sqrt{x}\sqrt{w}},\tag{21}$$

with $x=4s$. It means that we have a generating function for polynomials $\{f_n(x)\}_{n=0}^{\infty}$. Note that sn^{-1} is an analytic function within the unit circle. Hence, the only singularity of the right hand side of equation (21) is $w=0$. For our purposes, which are described below, this is an inconvenient form.

2 The measure of orthogonality

In the following section we will introduce the main idea and methods to obtain the measure of orthogonality of orthogonal polynomials in the case we are aware of their generating function. Then we show application of this process on polynomials $\{f_n(x)\}_{n=0}^{\infty}$.

2.1 The idea and methods

Collection of the following facts is mainly taken from [1] and [5].

Let μ denote the measure of orthogonality for orthogonal polynomials $\{P_n(x)\}_{n=0}^{\infty}$ corresponding to the Jacobi operator J (the unique self-adjoint extension of operator $\dot{J}$). Then

$$\mathrm{supp}(\mu)=\mathrm{spec}(J).$$

Thus we need to determine the spectrum of the operator J. We know that the spectrum of our Jacobi operators, determined by coefficients (13) and (14), is discrete due to [6]. Thus we know, that measure μ is of the form

$$\mu = \sum_{n=0}^{\infty} w_n \delta_{x_n}, \tag{22}$$

with δ_{x_n} being the Dirac measure centered in x_n.

We know that OG polynomials $\{P_n(x)\}_{n=0}^{\infty}$ satisfy the formal eigenvalue equation

$$\mathcal{J}\mathbf{P}(x) = x\mathbf{P}(x).$$

If it happens that $\mathbf{P}(x) \in \ell^2$, x is an eigenvalue of J with the eigenvector $\mathbf{P}(x)$. Assume that

$$P_n(x) = \frac{u(x)}{n^q} + o\left(\frac{1}{n}\right), \quad q \in (0,1). \tag{23}$$

Then $\{P_n(x)\}_{n=0}^{\infty} \in \ell^2$ only if $u(x) = 0$. This condition determines the spectrum of J. A device to obtain asymptotic behavior is called *the Darboux method*.

Theorem 1. Let f be a function with some isolated singularity not located in the origin of complex plane. Denote the distance of the singularity nearest to the origin by r. Let g be another function that obeys

1. g is holomorphic in $0 < |t| < r$,
2. $f - g$ is continuous in $0 < |t| \leq r$,
3. the coefficients b_n in Laurent expansion

$$g(t) = \sum_{n=-\infty}^{\infty} b_n t^n$$

have know asymptotic behavior.

Let

$$f(t) = \sum_{n=-\infty}^{\infty} a_n t^n.$$

Then

$$a_n = b_n + o\left(r^{-n}\right).$$

It remains to identify jumps w_n in (22). There exists a unique resolution of identity $\{E_\lambda\}_{\lambda\in\mathbb{R}}$ such that $J = \int_{\mathbb{R}} \lambda \mathrm{d}E_\lambda$. The orthogonal measure μ can be expressed as $\mu(\cdot) = \langle \mathbf{e}_0, E(\cdot)\mathbf{e}_0\rangle$. Then we can define *the Weyl m-function* corresponding to the operator J by

$$m_J(x) := \int_{\mathbb{R}} \frac{\mathrm{d}\mu(s)}{s - x},$$

which is the Borel transformation of the measure μ. It means that if we determine the LHS of the above equation somehow, we would be able to reconstruct the measure μ. According to *the Markov theorem*, see [3], one has

$$m_J(x) = \lim_{n\to\infty} \frac{p_{n-1}^{(1)}(x)}{p_n(x)}, \tag{24}$$

with $p_n^{(1)}(x)$ being associated polynomials. The Weyl m-function is meromorphic, thus there exist entire functions f, g such that $m_J(x) = \frac{f(x)}{g(x)}$. Jumps w_n then can be obtained as

$$w_n = \frac{f(x_n)}{g'(x_n)}. \tag{25}$$

The last thing we need to get is sequence $\left\{p_n^{(1)}(x)\right\}_{n=0}^{\infty}$.

Assume that matrix $\mathcal{J}^{(1)}$ arises from matrix $\mathcal{J}$ by deleting the first row and the first column of $\mathcal{J}$. This matrix is again a Jacobi matrix. So we can define another orthogonal polynomials $\left\{P_n^{(1)}(x)\right\}_{n=0}^{\infty}$ and monic orthogonal polynomials $\left\{p_n^{(1)}(x)\right\}_{n=0}^{\infty}$ corresponding to the matrix $\mathcal{J}^{(1)}$. For the monic polynomials we have

$$p_{n+1}^{(1)}(x) = (x - \beta_{n+1})p_n^{(1)}(x) - \alpha_n^2 p_{n-1}^{(1)}(x). \tag{26}$$

Polynomials $\left\{p_n^{(1)}(x)\right\}_{n=0}^{\infty}$ obey

$$\mathrm{Hn}^{(1)}(k^2, s; \alpha, \beta, \gamma, \delta; w) = \sum_{n=0}^{\infty} \frac{(-1)^n}{\nu_2\nu_3\ldots\nu_{n+1}} p_n^{(1)}(s + (\alpha+1)(\beta+1)k^2 - \delta k^2)w^n, \tag{27}$$

with $\mathrm{Hn}^{(1)}(k^2, s; \alpha, \beta, \gamma, \delta; w)$ beign given by

$$G(w) = w\mathrm{Hn}^{(1)}(k^2, s; \alpha, \beta, \gamma, \delta; w) \tag{28}$$

with $G(w)$ being the solution to the non-homogenous Heun equation

$$G''(w) + \left(\frac{\gamma}{w} - \frac{\delta}{1-w} - \frac{\epsilon k^2}{1-k^2 w}\right) G'(w) + \frac{s + \alpha\beta k^2 w + (\gamma+\epsilon)k^2}{w(1-w)(1-k^2w)} G(w) = \frac{\gamma}{w(1-w)(1-k^2w)}, \tag{29}$$

which is analytic in $|w| < 1$ and with initial data $G(0) = 0$ and $G'(0) = 1$.

Substitution for $w = \mathrm{sn}^2(z)$ and setting $v(z) := G(w)$ yield

$$v''(z) + \left((2\gamma-1)\frac{\mathrm{cn}(z)\mathrm{dn}(z)}{\mathrm{sn}(z)} - (2\delta-1)\frac{\mathrm{sn}(z)\mathrm{dn}(z)}{\mathrm{cn}(z)} - (2\epsilon-1)\frac{\mathrm{sn}(z)\mathrm{cn}(z)}{\mathrm{dn}(z)}\right) v'(z) + 4\left(s + (\gamma+\epsilon)k^2 + \alpha\beta k^2 \mathrm{sn}^2(z)\right) v(z) = 4\gamma, \tag{30}$$

with initial data $v(0) = v'(0) = 0$ and $v''(0) = 2$.

Assume that $v_1(z)$ and $v_2(z)$ are two linearly independent solutions to the homogenous Heun equation (30). The the solution $v(z)$ to the inhomogenous Heun solution (30) according to [**?**] reads

$$v(z) = C \int_0^z \frac{v_1(z)v_2(t) - v_1(t)v_2(z)}{\mathrm{sn}^{1-2\gamma}(t)\mathrm{cn}^{1-2\delta}(t)\mathrm{dn}^{1-2\epsilon}(t)}\mathrm{d}t \tag{31}$$

with C being a constant independent on z which need to be set in correspondence with the initial data.

2.2 The second comeback to the motivation

As claimed before, expression (21) is inconvenient for our purpose. In particular, it is not in a suitable form for using the Darboux method. Derivative of the both sides of equation (21) yields

$$\sum_{n=0}^{\infty} \frac{f(-x)}{(2n)!} = \frac{\cos\left(\sqrt{x}\ \mathrm{sn}^{-1}(\sqrt{w})\right)}{\sqrt{1-w}\sqrt{1-k^2w}}. \tag{32}$$

Recall that $x = 4s$. The RHS of the preceding equation is suitable for using the Darboux method, since the nearest singularity to the origin is $w = 1$, and is the generating function for polynomials $\{f_n(x)\}_{n=0}^{\infty}$ at the same time. Note that

$$\mathrm{Hn}\left(k^2, s - \frac{1}{4} - \frac{k^2}{4}; 1, \frac{3}{2}, \frac{1}{2}, \frac{3}{2}; w\right) = \frac{\cos\left(\sqrt{x}\ \mathrm{sn}^{-1}(\sqrt{w})\right)}{\sqrt{1-w}\sqrt{1-k^2w}}. \tag{33}$$

This result could be again obtained by symmetries of the Heun equation, see [9]. Using the Darboux method on equation (32) yields asymptotic behavior

$$\begin{aligned}\frac{f_n(-x)}{(2n)!} &= \frac{\cos(\sqrt{x}K)}{\sqrt{1-k^2}}(-1)^n\binom{-1/2}{n} + o(1)\\ &\sim \frac{(-1)^n}{\sqrt{\pi n}}\frac{\cos(\sqrt{x}K)}{\sqrt{1-k^2}} + o(1), \quad n \to \infty.\end{aligned}$$

Therefore eigenvalues of the corresponding Jacobi operator read

$$s_n = \frac{(2n+1)^2\pi^2}{16K^2} - \frac{1}{4} + \frac{5}{4}k^2. \tag{34}$$

From (33) one has

$$\begin{aligned}\mathrm{Hn}^{(1)}\left(k^2, s; 1, \frac{3}{2}, \frac{1}{2}, \frac{3}{2}; w\right) = &\frac{\mathrm{sn}^{-1}\sqrt{w}}{\sqrt{s + \frac{1}{4} + \frac{9}{4}k^2}w\sqrt{1-w}\sqrt{1-k^2w}}\times\\ &\times \int_0^1 \sin\left(2\sqrt{s + \frac{1}{4} + \frac{9}{4}k^2}(1-\tau)\sqrt{w}\right)\mathrm{cn}(\tau z)\mathrm{dn}(\tau z)\mathrm{d}\tau.\end{aligned}$$

Let us denote

$$\mathfrak{X}(s) := \frac{K}{\sqrt{s+\frac{1}{4}+\frac{9}{4}k^2}\sqrt{1-k^2}} \int_0^1 \sin\left(2K\sqrt{s+\frac{1}{4}+\frac{9}{4}k^2}(1-\tau)\right) \mathrm{cn}(K\tau)\mathrm{dn}(K\tau)\mathrm{d}\tau.$$

Thus

$$\frac{p_{n-1}^{(1)}(s)}{p_n(s)} \sim -2\frac{\sqrt{1-k^2}\mathfrak{X}\left(s-\frac{7}{2}k^2\right)}{\cos\left(K\sqrt{4s+1-5k^2}\right)}\sqrt{\frac{n}{n-1}}, \quad \text{for } n\to\infty.$$

It means

$$m_J(s) = \lim_{n\to\infty}\frac{p_{n-1}^{(1)}(s)}{p_n(s)} = -2\frac{\sqrt{1-k^2}\mathfrak{X}\left(s-\frac{7}{2}k^2\right)}{\cos\left(K\sqrt{4s+1-5k^2}\right)}. \tag{35}$$

Recall that w_n are given by (25) with

$$g(s) = \cos\left(K\sqrt{4s+1-5k^2}\right), \qquad g'(s) = -\frac{2K\sin\left(K\sqrt{4s+1-5k^2}\right)}{\sqrt{4s+1-5k^2}}$$

and

$$f(s) = \frac{-2K}{\sqrt{s+\frac{1}{4}-\frac{5}{4}k^2}} \int_0^1 \sin\left(2K\sqrt{s+\frac{1}{4}-\frac{5}{4}k^2}(1-t)\right) \mathrm{cn}(Kt)\mathrm{dn}(Kt)\mathrm{d}t.$$

After a routine manipulation one arrives to

$$w_n = \frac{(2n+1)\pi^2}{K^2k}\frac{q^{n+\frac{1}{2}}}{1-q^{2n+1}}. \tag{36}$$

We can see that (34) and (36) agree with (2).

3 Interim results

We desire to exhaust all options of new families of orthogonal polynomials constructed from the level one solution. Our starting point will be the following theorem from [10] in which G. Valent claims how do those solutions look like.

Theorem 2. The level 1 solutions of equation (6) have the following forms with $Z(\omega) = \frac{\Theta'(\omega)}{\Theta(\omega)}$.

	$m_0 = 1$	$m_1 = 1$	$m_2 = 1$	$m_3 = 1$
solution	$y_1(z) = e^{z\mathbb{Z}(\omega)}\frac{H(z-\omega)}{\Theta(z)}$	$y_2(z) = e^{z\mathbb{Z}(\omega)}\frac{\Theta(z-\omega)}{\Theta(z)}$	$y_3(z) = e^{z\mathbb{Z}(\omega)}\frac{\Theta_1(z-\omega)}{\Theta(z)}$	$y_4(z) = e^{z\mathbb{Z}(\omega)}\frac{H_1(z-\omega)}{\Theta(z)}$
condition	$\mathrm{dn}^2(\omega) = 4s-k^2$	$\mathrm{dn}^2(\omega) = 4s+1$	$\mathrm{dn}^2(\omega) = 4s+1-k^2$	$\mathrm{dn}^2(\omega) = 4s+1-k^2$

From now on, the Heun equation will be represented by vectors

$$(s;\alpha,\beta,\gamma,\delta,\epsilon)\in\mathbb{C}^6.$$

Talking about the solutions of the Heun equation y, if the dependence on z is emphasized, i.e. $y(z)$, the solution is meant after the transformation $w=\mathrm{sn}^2(z)$ of the Heun equation. Assume the Heun equation $(s;\alpha,\beta,\gamma,\delta,\epsilon)$ with a solution $y(z)$ and function $u(z)$. In the left column of the below chart we set a relationship between $y(z)$ and $u(z)$. In the right column there is a new Heun equation for which $u(z)$ is a solution.

$y(z)=\mathrm{cn}^{2(1-\delta)}(z)u(z)$	$(s+\gamma(\delta-1);\alpha-\delta+1,\beta-\delta+1,\gamma,2-\delta,\epsilon)$
$y(z)=\mathrm{dn}^{2(1-\epsilon)}(z)u(z)$	$(s+k^2\gamma(\epsilon-1);-\alpha+\gamma+\delta,-\beta+\gamma+\delta,\gamma,\delta,2-\epsilon)$
$y(z)=\mathrm{sn}^{2(1-\gamma)}(z)u(z)$	$(s+(\gamma-1)(\delta+k^2\epsilon);\beta-\gamma+1,\alpha-\gamma+1,2-\gamma,\delta,\epsilon)$

For the later purposes let us define (non-linear) operators $T_1,T_2,T_3:\mathbb{C}^6\to\mathbb{C}^6$: $(s;\alpha,\beta,\gamma,\delta,\epsilon)\to(\tilde{s};\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta},\tilde{\epsilon})$ where output corresponds to the vector in the first, respectively the second, respectively the third row of the above chart.

Having the level 1 solutions y_1,y_2,y_3 and y_4 and transformations T_1,T_2 and T_3, a natural question arises: How many new families we can obtain?

Denote $G:=\{\mathrm{id},T_1,T_2,T_3,T_1T_2,T_2T_3,T_1T_3,T_1T_2T_3\}$.

Proposition 3. Set G together with composition of the mappings and an inverse mapping forms an Abelian group of the eighth order.

Proof. By the direct computation one obtains

$$T_i^2=\mathrm{id},\quad i\in\{1,2,3\};\tag{37}$$
$$T_iT_j=T_jT_i,\quad i,j\in\{1,2,3\}.\tag{38}$$

Equation (38) yields

$$T_iT_jT_k=T_{\pi(i)}T_{\pi(j)}T_{\pi(k)},\quad i,j,k\in\{1,2,3\},\quad \pi\in S_3.\tag{39}$$

It is readily seen that for any $A\in G$, $A^{-1}=A$. Thus the inversion applied on elements of G does not yield any other transformations. Similarly, for any $A,B\in G$, $AB\in G$. Thus, the composition does not provide any new transformations as well. □

And so the answer is, that there are no more then 32 new families of orthogonal polynomials which can be constructed from the level 1 solution of the Heun equation. Assume two families $\{p_n(x)\},\{\tilde{p}_n(x)\}$ with coefficients α_n,β_n, resp. $\tilde{\alpha}_n,\tilde{\beta}_n$.

- If $\alpha_n=\tilde{\alpha}_n$ and $\beta_n-\tilde{\beta}_n=const.$, then families $\{p_n(x)\}$ and $\{\tilde{p}_n(x)\}$ coincides.
- If $\alpha_0=\beta_0$, then $p_n(x)=xq_n(x)$ with family $\{q_n(x)\}$ given by

$$q_{n+1}(x)=(x-\beta_{n+1})q_n(x)-\alpha_n^2q_{n-1}(x),\quad n\in\mathbb{N},\quad q_{-1}(x)=0,\quad q_0(x)=1.\tag{40}$$

Using the above items allows decrease the number of new families to 12. For the brevity we don't list all 32 potentially new families. It follows from the specific forms for different families. We do not list them for brevity.

Let us ilustrate results about the measure of orthogonality on the example of one particular family. According to theorem 2, function $y_1(z)$ is a solution to the Heun equation $(s; -\frac{1}{2}, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. Using transformation T_3T_1 one gets the Heun equation $\left(s - 2 - \frac{3k^2}{4}; 2, \frac{3}{2}, \frac{5}{2}, \frac{3}{2}, \frac{1}{2}\right)$ with the solution

$$y(z) = e^{zZ(\omega)}\frac{\Theta(z-\omega)}{\mathrm{sn}^3(z)H_1(z)}, \quad 4s+1 = \mathrm{dn}^2(\omega).$$

It is easy to see that $y(-z)$ is the solution of the same Heun equation and that $y(z)$ and $y(-z)$ are linearly independent. Thus, the respective Heun function, the generating function for orthogonal polynomials $\{p_n(x)\}_{n=0}^{\infty}$, can be written as a linear combination of these two solutions. After a rutine manipulation with the Jacobi elliptic functions and theta functions we have

$$\mathrm{Hn}\left(k^2, s-\frac{1}{4}; 0, \frac{3}{2}, \frac{1}{2}, \frac{3}{2}, \mathrm{sn}^2(z)\right) = \frac{\Theta(0)}{2H(\omega)\Theta(z)\mathrm{cn}(z)}\left(e^{-zZ(\omega)}H(z+\omega) - e^{zZ(\omega)}H(z-\omega)\right).$$

Repeating the same procedure as in the case of the family $\{f_n(x)\}_{n=0}^{\infty}$, one gets

$$\frac{p_n(x)}{(2n+2)!} = -\frac{1}{4^n k^2\sqrt{\pi n}}\frac{\sinh(KZ(\omega))}{\mathrm{sn}(\omega)\mathrm{cn}(\omega)} + O\left(\frac{1}{n4^n}\right), \quad n\to\infty.$$

After an analysis of zeros of the function $\frac{\sinh(KZ(\omega))}{\mathrm{sn}(\omega)\mathrm{cn}(\omega)}$ one arrives to the following conditions for eigenvalues of the corresponding Jacobi operator J

$$4x_n = \mathrm{dn}^2(\omega_n), \quad \omega_n = \mathrm{i}v_n, \quad \text{for } v_n\in\mathbb{R}, \quad \frac{KH_1(v_n)}{H_1(v_n)} + \frac{\pi v_n}{2K'} = 2n\pi.$$

For the Weyl m-function of the Jacobi operator J one has

$$m_J(x) = \lim_{n\to\infty}\frac{p_{n-1}^{(1)}(x)}{q_n(x)} = -\frac{2k^2}{15}\frac{\mathrm{sn}(\omega)\mathrm{cn}(\omega)\mathfrak{Y}(\Omega;K)}{\sinh(KZ(\omega)},$$

with

$$\mathfrak{Y}(\Omega;z) = \frac{C(\Omega)}{\mathrm{sn}^5(z)}\int_0^z (y_1(z)y_1(-t) + y_1(t)y_1(-z))\mathrm{sn}(t)\mathrm{cn}(t)\mathrm{d}t,$$

here $C(\Omega)$ is, for now, not further specified constant depending on Ω and $4x = \mathrm{dn}(\Omega)$.

References

[1] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis*, Oliver and Boyd, Edinburgh, 1965.

[2] M. Abramowitz, I. A. Stegun, *Handbook of mathematical functions*, National Bureau of Standards, 1964.

[3] C. Berg, *Markov's theorem revisited*, J. Approx. Theory 78 (1994) 260–275.

[4] L. Carlitz, *Some orthogonal polynomials related to elliptic functions*, Duke Mathematical Journal 27 (4), 443 – 459 (1960).

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

[6] J. Janas, S. Naboko, *Multitreshold Spectral Phase Transitions for a Class of Jacobi Matrices*, Operator Theory: Advances and Applications Vol. 124 (2001) 267–285.

[7] D. Lawden *Elliptic functions and Applications*, Springer-Verlag New York Inc., New York, 1989.

[8] F. W. J. Olver *Asymptotics and Special Functions*, New York: Academic Press, New York, 1997.

[9] A. Ronveaux, *Heun's Differential Equation*, Oxford University Press, Oxford, 1995.

[10] G. Valent, *Heun functions versus elliptic functions*, Difference equation, special functions and orthogonal polynomials, Eds. S. Elaydi et al., World Scientific Hackensack (2007), 664–668.

Discrete Orthogonality of Orbit Functions Arising from the Root System C_2*

Vojtěch Teska
vojtech.teska@fjfi.cvut.cz

study programme: Mathematical Engineering
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

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

Lenka Motlochová, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. In the root system C_2, there are four classical Weyl-group-invariant lattices: the root lattice, the coroot lattice, the weight lattice and the coweight lattice. All four can be used to construct a generalized affine Weyl group via their semidirect product with the Weyl group of the root system in question. Generalized affine Weyl groups admit four classes of functions invariant with respect to their action on their arguments, called orbit functions. Discretizing the domain of the orbit functions and restricting to the fundamental domain of the action of the generalized affine Weyl group, one can find orthogonality relations which can serve as the basis of Fourier-type analysis of discrete functions.

Keywords: root system C_2, orbit function, Weyl group, invariant lattice

Abstrakt. V kořenovém systému C_2 existují čtyři klasické mříže invariantní vůči Weylově grupě: kořenová mříž, mříž duálních kořenů, váhová mříž a mříž duálních vah. Všechny čtyři mohou být použity ke konstrukci zobecněné afinní Weylovy grupy pomocí polopřímého součinu s Weylovou grupou příslušnou tomuto kořenovému systému. Zobecněné afinní Weylovy grupy připouštějí čtyři třídy funkcí invariantních vůči jejich akcím na jejich argumenty, zvané orbitové funkce. Diskretizací a omezením na fundamentální oblast akce zobecněné afinní Weylovy grupy lze dokázat ortogonalitu orbitových funkcí, která může sloužit jako základ Fourierovy analýzy diskrétních funkcí.

Klíčová slova: kořenový systém C_2, orbitová funkce, Weylova grupa, invariantní mříž

1 Introduction

The purpose of this article is to give an exhaustive description of orthogonality relations of discretized orbit functions arising from the four classical Weyl-group-invariant lattices constructed from the root system C_2 [6]. Orbit functions are complex functions labelled by two-dimensional real vectors whose arguments are also elements of $\mathbb{R}^2$. They take form

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

of (anti)-symmetrized sums of exponents of scalar products of their labels and arguments over the Weyl group of the corresponding root system [3, 4, 7, 8]. The invariance of orbit functions with respect to translations and the action of the Weyl group allows us to restrict the study of orbit functions to the fundamental domain of the corresponding generalized affine Weyl group. A finite fragment of a rescaled lattice situated inside the fundamental domain serves as the set of sampling points on which the orbit functions are discretized. The finite-dimensional vector space of complex functions on this fragment admits an orthogonal basis of orbit functions, labelled by elements of a different Weyl-group-invariant lattice [1, 3–5].

This article is organized as follows: in Section 2, the necessary facts concerning the root system C_2 are recalled. Section 3 contains the description of the four classical Weyl-group-invariant lattices: the root lattice, the coroot lattice, the weight lattice and the coweight lattice. In Section 4, the generalized affine Weyl groups are defined and their fundamental domains are depicted explicitly. The following section is dedicated to the definition and properties of orbit functions and the construction of the sampling and label grids stemming from different lattices. Section 6 is divided into four subsections, each discussing related cases of discrete orthogonality. Comments and follow-up discussion are contained in the conclusion.

2 Root system C_2

The root system C_2, which will be denoted Π in this article, is a crystallographic root system in $\mathbb{R}^2$ determined by its Cartan matrix

$$C = \begin{pmatrix} 2 & -1 \\ -2 & 2 \end{pmatrix}. \tag{1}$$

The elements of C are defined as

$$C_{ij} = \frac{2(\alpha_i, \alpha_j)}{(\alpha_j, \alpha_j)} \quad i, j \in \{1, 2\} \tag{2}$$

where α_1, α_2 are the so-called simple roots [6]. This matrix uniquely determines the ratio of the squared lengths of α_1 and α_2 and the angle between them, so to determine Π uniquely, we adopt the additional convention for the length of the long root

$$(\alpha_2, \alpha_2) = 2, \tag{3}$$

so we arrive to the conclusion that

$$||\alpha_1|| = 1, \quad ||\alpha_2|| = \sqrt{2} \tag{4}$$

and that the angle between α_1 and α_2 is equal to $\frac{3}{4}\pi$.

Π is by definition invariant with respect to its corresponding Weyl group W, which is generated by the reflections r_α where α is any element of Π. The reflection r_α is the linear map

$$r_\alpha \cdot x = x - \frac{2(x, \alpha)}{(\alpha, \alpha)}\alpha \tag{5}$$

where $x \in \mathbb{R}^2$. For future convenience, denote $r_{1,2} := r_{\alpha_{1,2}}$. Using the fact that r_1, r_2 generate the Weyl group [6], we determine that $|W| = 8$. The entire Π is obtained from $\{\alpha_1, \alpha_2\}$ by the action of W. Finally, let us remark that

$$\frac{2(\alpha,\beta)}{(\beta,\beta)} \in \mathbb{Z} \quad \text{for all} \quad \alpha, \beta \in \Pi. \tag{6}$$

3 W-invariant lattices of C_2

Given $\alpha \in \Pi$, we define the coroot $\alpha^\vee$ by

$$\alpha^\vee := \frac{2\alpha}{(\alpha,\alpha)} \tag{7}$$

for all $\alpha \in \Pi$. The vectors $\alpha_1^\vee, \alpha_2^\vee$ are called the simple coroots. The fundamental weights are defined by the relation

$$(\alpha_i^\vee, \omega_j) = \delta_{ij} \quad i,j = 1,2. \tag{8}$$

Analogously, the fundamental coweights are defined by

$$(\omega_j^\vee, \alpha_j) = \delta_{ij} \quad i,j = 1,2. \tag{9}$$

We see that the coroots as well as the fundamental weights and fundamental coweights all form a basis of $\mathbb{R}^2$. Expressing these vectors in terms of the simple roots α_1, α_2 using the equations (7), (8) and (9), we obtain:

$$\alpha_1^\vee = 2\alpha_1, \qquad \omega_1 = \alpha_1 + \frac{1}{2}\alpha_2, \qquad \omega_1^\vee = 2\alpha_1 + \alpha_2, \tag{10}$$

$$\alpha_2^\vee = \alpha_2, \qquad \omega_2 = \alpha_1 + \alpha_2, \qquad \omega_2^\vee = \alpha_1 + \alpha_2, \tag{11}$$

as depicted in Figure 1.

The root lattice is defined as the $\mathbb{Z}$-span of the simple roots:

$$Q = \mathbb{Z}\alpha_1 + \mathbb{Z}\alpha_2, \tag{12}$$

the coroots give rise to the coroot lattice $Q^\vee$ which is defined as

$$Q^\vee = \mathbb{Z}\alpha_1^\vee + \mathbb{Z}\alpha_2^\vee, \tag{13}$$

the weights generate the weight lattice P

$$P = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2, \tag{14}$$

and the coweights span the coweight lattice $P^\vee$

$$P^\vee = \mathbb{Z}\omega_1^\vee + \mathbb{Z}\omega_2^\vee. \tag{15}$$

A lattice $A \subseteq \mathbb{R}^2$ is W-invariant, if

$$w \cdot A \subseteq A \quad \text{for all} \quad w \in W. \tag{16}$$

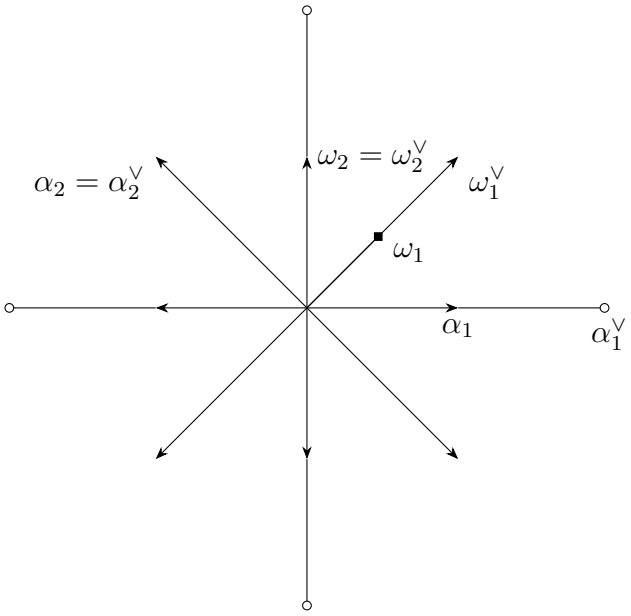


Figure 1: The root system C_2 with its coroots, fundamental weights and fundamental coweights. The roots have arrow tips and coroots not belonging to C_2 have circle tips. The fundamental weight ω_1 has a square tip.

We can check using the crystallographic condition (6) that Q and $Q^\vee$ are both W-invariant. Given a lattice $A \subseteq \mathbb{R}^2$, we define its dual lattice [6] $A^\perp$ by the requirement

$$A^\perp = \left\{ v \in \mathbb{R}^2 \mid (v, a) \in \mathbb{Z}, \forall a \in A \right\}. \tag{17}$$

Equations (8) and (9) give us $P = (Q^\vee)^\perp$ and $P^\vee = Q^\perp$, respectively. If a lattice A is W-invariant, then its dual $A^\perp$ is W-invariant as well, hence both P and $P^\vee$ are W-invariant. Furthermore, let us note that (6) gives us the inclusion $Q \subseteq P$ which also dualizes to $Q^\vee \subseteq P^\vee$. The inclusions $Q^\vee \subseteq Q$ and $P^\vee \subseteq P$ are obvious from the definitions (7) and (9). Finally, the explicit forms of the vectors give us $Q = P^\vee$ and $2P = Q^\vee$. Summarizing our results for the root system C_2, we have:

$$2P = Q^\vee \subseteq Q = P^\vee \subseteq P. \tag{18}$$

4 Fundamental domains of group actions

Let us consider the group G acting on $\mathbb{R}^2$ generated by all reflections $r_\alpha, \alpha \in \Pi$ and all shifts $T(a)$ where $a \in A$ is an element of some W-invariant lattice A. The subgroup generated solely by the shifts is isomorphic to the lattice A and the subgroup generated by $r_\alpha, \alpha \in \Pi$ is W. Since $wT(a) = T(w \cdot a)w$, we have $G = AW$, a similar calculation shows that $wT(a)w^{-1} = T(w \cdot a)$, so A is normal in G and clearly $W \cap A = \{\mathrm{id}_{\mathbb{R}^2}\}$, hence $G = A \rtimes W$. Any group of transformations of $\mathbb{R}^2$ obtained in this manner is called the generalized affine Weyl group. The fundamental domain of a generalized affine Weyl group is a subset F_A of $\mathbb{R}^2$ such that

1. $(A \rtimes W)F_A = \mathbb{R}^2$

2. F_A contains at most one point from every orbit of $A \rtimes W$.

The fundamental domains of the generalized affine Weyl groups using the lattices $Q, Q^\vee, P$ and $P^\vee$ can be found in the following manner: realize that a convenient set of representants of $\mathbb{R}^2/A$ is formed by the set $S_A = \{a_1 v_1 + a_2 v_2 \mid a_1, a_2 \in I\} \cup \{0\}$. The vectors $v_{1,2}$ are $\alpha_{1,2}$ for Q, $\alpha_{1,2}^\vee$ for $Q^\vee$, $\omega_{1,2}$ for P, $\omega_{1,2}^\vee$ for $P^\vee$ and I is the interval $(0,1]$ for the lattices $Q, Q^\vee$ and $[0,1)$ for the lattices $P, P^\vee$. At this point, it suffices to find a set F_A such that there exist $w \in W, a \in A$ and $y \in F_A$ satisfying the equation $x = w \cdot y + a$ for every $x \in S_A$. The results are

$$F_A = \left\{ y_1^A q_1^A \omega_1 + y_2^A q_2^A \omega_2 \mid y_0^A + y_1^A + y_1^A = 1 \right\}, \tag{19}$$

where

$$q_1^A = \begin{cases} 1 & A = Q, Q^\vee, P^\vee \\ \frac{1}{2} & A = P \end{cases} \qquad q_2^A = \begin{cases} 1 & A = Q^\vee \\ \frac{1}{2} & A = Q, P, P^\vee. \end{cases} \tag{20}$$

The fundamental domains F_A are depicted in Figure 2. The counting functions

$$\varepsilon_A(s) : F_A \to \mathbb{N}, \quad \varepsilon_A(s) = |W \cdot s|\,; \tag{21}$$

$$h_A^M(\lambda) : MF_A \to \mathbb{N}, \quad h_A^M(\lambda) = \left| \mathrm{Stab}\left(\frac{\lambda}{M}\right) \right|; \tag{22}$$

where the action of W is considered on the torus $\mathbb{R}^2/A$, play a crucial role in the following sections. The values of ε_A for the classical invariant lattices are as follows:

$$\varepsilon_Q(s) = \begin{cases} 8 & s \in \mathrm{int}(F_Q) \\ 4 & s \in \partial F_Q \setminus \{0, \omega_1, \frac{1}{2}\omega_2, \} \\ 2 & s = \frac{1}{2}\omega_2 \\ 1 & s \in \{0, \omega_1\} \end{cases} \qquad \varepsilon_{Q^\vee}(s) = \begin{cases} 8 & s \in \mathrm{int}(F_{Q^\vee}) \\ 4 & s \in \partial F_{Q^\vee} \setminus \{0, \omega_1, \omega_2\} \\ 2 & s = \omega_1 \\ 1 & s \in \{0, \omega_2\} \end{cases} \tag{23}$$

$$\varepsilon_P(s) = \begin{cases} 8 & s \in \mathrm{int}(F_P) \\ 4 & s \in \partial F_P \setminus \{0, \frac{1}{2}\omega_1, \frac{1}{2}\omega_2\} \\ 2 & s = \frac{1}{2}\omega_1 \\ 1 & s \in \left\{0, \frac{1}{2}\omega_2\right\} \end{cases} \qquad \varepsilon_{P^\vee}(s) = \varepsilon_Q(s) \quad s \in F_{P^\vee} = F_Q, \tag{24}$$

where $\mathrm{int}(\cdot)$ denotes the interior of a set and $\partial(\cdot)$ denotes its boundary. These values are obtained by using all $w \in W$ on the fundamental domain, shifting the result to S_A and counting the occurrences of each point. The numbers h_A^M are obtained using the values of ε_A and the orbit-stabilizer theorem.

5 Orbit functions

Orbit functions are complex functions respecting the symmetries of the generalized affine Weyl group. First, note that there are four possible homomorphisms of W into the multiplicative cyclic group $\mathbb{Z}_2 = \{-1, 1\}$, given by

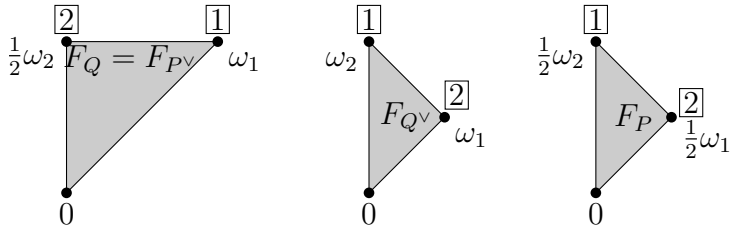


Figure 2: Fundamental domains of generalized affine Weyl groups of the root system C_2. The values of the function ε_A in the corner points are given by the numbers in the boxes.

$$\mathbb{1}(r_i) = 1, \qquad \sigma^e(r_i) = -1, \qquad \sigma^s(r_i) = \begin{cases} -1 & i = 1 \\ 1 & i = 2 \end{cases}, \qquad \sigma^l(r_i) = \begin{cases} 1 & i = 1 \\ -1 & i = 2 \end{cases}, \tag{25}$$

as proven in [9].

The orbit function $\varphi_b^\sigma : \mathbb{R}^2 \to \mathbb{C}$, where σ is any of the multiplicative homomorphisms from (25) and $b \in \mathbb{R}^2$ is its label, is defined as

$$\varphi_b^\sigma(x) = \sum_{w \in W} \sigma(w) e^{2\pi \mathrm{i}(w \cdot b, x)} \tag{26}$$

for any $x \in \mathbb{R}^2$. The functions $\varphi_b^{\mathbb{1}}$ are called C-functions, the functions $\varphi_b^{\sigma^e}$ are called S-functions and the functions $\varphi_b^{\sigma^s}$ and $\varphi_b^{\sigma^l}$ are called S^s-functions and S^l-functions respectively [3, 4].

Let us summarize their pertinent properties:

$$\varphi_b^\sigma(w \cdot x) = \sigma(w)\varphi_b^\sigma(x), \tag{27}$$
$$\varphi_{w \cdot b}^\sigma(x) = \sigma(w)\varphi_b^\sigma(x), \tag{28}$$

for any $w \in W$ and $b, x \in \mathbb{R}^2$. If B is a W-invariant lattice, then for any $b \in B, b^\perp \in B^\perp$ and $x \in \mathbb{R}^2$ it holds that

$$\varphi_b^\sigma(x + b^\perp) = \varphi_b^\sigma(x), \tag{29}$$

from which it follows that φ_b^σ is well-defined on $\mathbb{R}^2/B^\perp$. If A is a W-invariant lattice, then for any $M \in \mathbb{N}, a \in A, b \in B$ and $a^\perp \in A^\perp$ it holds that

$$\varphi_{b+Ma^\perp}^\sigma \left(\frac{1}{M} a \right) = \varphi_b^\sigma \left(a^\perp \right), \tag{30}$$

so we can regard orbit functions as complex functions on $B/MA^\perp \times \frac{1}{M} A/B^\perp$ if the relation

$$A^\perp \subseteq B \Leftrightarrow B^\perp \subseteq A \tag{31}$$

holds.

Define the sets $F_A^\sigma \subseteq F_A$ as follows:

$$F_A^\sigma = \left\{ y_1^{A,\sigma} q_1^A \omega_1 + y_2^{A,\sigma} q_2^A \omega_2 \mid y_0^{A,\sigma} + y_1^{A,\sigma} + y_1^{A,\sigma} = 1 \right\}, \tag{32}$$

where $q_{1,2}$ are given by (20). The numbers $y_i^{A,\sigma}$ have the following properties: $y_i^{A,\mathbb{1}} \geq 0$ for $i = 0, 1, 2$; $y_i^{A,\sigma^e} > 0$ for $i = 0, 1, 2$ and

$$y_0^{A,\sigma^s} \begin{cases} \geq 0 & A = Q^\vee, P \\ > 0 & A = Q, P^\vee \end{cases} \qquad y_1^{A,\sigma^s} > 0 \quad \text{for all } A, \qquad y_2^{A,\sigma^s} \geq 0 \quad \text{for all } A; \tag{33}$$

$$y_0^{A,\sigma^l} \begin{cases} \geq 0 & A = Q, P^\vee \\ > 0 & A = Q^\vee, P \end{cases} \qquad y_1^{A,\sigma^l} \geq 0 \quad \text{for all } A, \qquad y_2^{A,\sigma^l} > 0 \quad \text{for all } A. \tag{34}$$

One can check on a case-by-case basis that the sets F_A^σ were chosen so that for every $y \in \frac{1}{M}A \cap \left(F_{B^\perp} \setminus F_{B^\perp}^\sigma\right)$ there exist $b^\perp \in B^\perp, r \in W$ such that $\sigma(r) = -1$ and $y = r \cdot y + b^\perp$, which implies $\varphi_b^\sigma(y) = 0$ for all $b \in B$. Similarly, for all $b \in B \cap \left(MF_{A^\perp} \setminus MF_{A^\perp}^\sigma\right)$ there exist $a^\perp \in A^\perp, r \in W$ such that $\sigma(r) = -1$ and $b = r \cdot b + Ma^\perp$, hence $\varphi_b^\sigma\left(\frac{1}{M}a\right) = 0$ for every $a \in A$.

We define the set of sampled points as

$$F_{B,A}^{\sigma,M} = \frac{1}{M}A \cap F_{B^\perp}^\sigma \tag{35}$$

and the set of labels as

$$\Lambda_{B,A}^{\sigma,M} = B \cap MF_{A^\perp}^\sigma. \tag{36}$$

The sets $F_{B,A}^{\sigma,M}$ and $\Lambda_{B,A}^{\sigma,M}$ are the minimal sets of points and labels respectively in which the corresponding orbit function is non-zero. Moreover, due to the symmetry relations (27) and (29), the values of φ_b^σ on the entire lattice $\frac{1}{M}A$ can be reconstructed from the values on $F_{B,A}^{\sigma,M}$.

The relationships between the lattices $Q, Q^\vee, P$ and $P^\vee$ given by (18) and the requirement (31) only allow six different pairs of lattices from which to construct respective label and point sets. These are: $(P, Q^\vee)$, $(Q^\vee, P)$, $(Q, P^\vee)$, $(P, P^\vee)$, $(P^\vee, P)$ and (P, P) where the first position in the pair determines the elements of the set of labels and the second determines the elements of the set of sampled points.

6 Orthogonality relations

We introduce the scalar product of discrete functions f, g on the point set $F_{B,A}^{\sigma,M}$ by

$$\langle f, g\rangle_{B,A}^{\sigma,M} = \sum_{s \in F_{B,A}^{\sigma,M}} \varepsilon_{B^\perp}(s) f(s) \overline{g(s)}, \tag{37}$$

where the number $\varepsilon_{B^\perp}$ is defined by (21).

6.1 The cases $(P, Q^\vee)$ and $(Q^\vee, P)$

The case $(P, Q^\vee)$ has already been discussed in [1]. It turns out that that the grids $F_{P,Q^\vee}^{\sigma,M}$ and $\Lambda_{P,Q^\vee}^{\sigma,M}$ have the same cardinality and that

$$\langle \varphi_\lambda, \varphi_\nu\rangle_{P,Q^\vee}^{\sigma,M} = 8M^2 h_P^M(\lambda)\delta_{\lambda,\nu} \tag{38}$$

for any $\lambda, \nu \in \Lambda^{\sigma,M}_{P,Q^\vee}$, hence the set of orbit functions indexed by elements of $\Lambda^{\sigma,M}_{P,Q^\vee}$ forms an orthogonal basis of the space of complex functions on $F^{\sigma,M}_{P,Q^\vee}$.

Let us turn our attention to the dual case $(Q^\vee, P)$. The relation (18) tells us that there is a one-to-one correspondence between the sets $\Lambda^{\sigma,M}_{Q^\vee,P}$, $\Lambda^{\sigma,M}_{P,Q^\vee}$ and another one-to-one correspondence between $F^{\sigma,M}_{Q^\vee,P}$ and $F^{\sigma,M}_{P,Q^\vee}$. If

$$\frac{2}{M}P/2P \ni s \leftrightarrow \tilde{s} \in \frac{1}{M}P/P, \tag{39}$$

then $\varepsilon_{Q^\vee}(s) = \varepsilon_P(\tilde{s})$ and similarly for

$$P/MP \ni \lambda \leftrightarrow \tilde{\lambda} \in 2P/2MP \tag{40}$$

it holds that $h^M_P(\lambda) = h^M_{Q^\vee}(\tilde{\lambda})$, hence it follows that

$$\langle \varphi_{\tilde{\lambda}}, \varphi_{\tilde{\nu}} \rangle^{\sigma,M}_{Q^\vee,P} = 8M^2 h^M_P(\tilde{\lambda}) \delta_{\tilde{\lambda},\tilde{\nu}}, \tag{41}$$

for any $\tilde{\lambda}, \tilde{\nu} \in \Lambda^{\sigma,M}_{Q^\vee,P}$, so the orbit functions indexed by $\Lambda^{\sigma,M}_{Q^\vee,P}$ form an orthogonal basis of the space of complex functions on $F^{\sigma,M}_{Q^\vee,P}$.

6.2 The case $(Q, P^\vee)$

Since $Q = P^\vee$, so $F_Q = F_{P^\vee}$, this case can also be thought of as having labels from $\Lambda^{\sigma,M}_{Q,Q}$ and points from $F^{\sigma,M}_{Q,Q}$. Clearly, $\Lambda^{\sigma,M}_{Q,Q} = MF^{\sigma,M}_{Q,Q}$, so the sets $\Lambda^{\sigma,M}_{Q,Q}$ and $F^{\sigma,M}_{Q,Q}$ must contain the same number of points. We can calculate the magnitudes of point sets to be

$$\left|F^{\mathbb{1},M}_{Q,Q}\right| = \begin{cases} \frac{M^2}{8} + \frac{3}{4}M + 1 & M \text{ even} \\ \frac{M^2}{8} + \frac{1}{2}M + \frac{3}{8} & M \text{ odd} \end{cases}, \quad \left|F^{\sigma^e,M}_{Q,Q}\right| = \begin{cases} \frac{M^2}{8} - \frac{3}{4}M + 1 & M \text{ even} \\ \frac{M^2}{8} - \frac{1}{2}M + \frac{3}{8} & M \text{ odd} \end{cases}, \tag{42}$$

$$\left|F^{\sigma^s,M}_{Q,Q}\right| = \begin{cases} \frac{M^2}{8} - \frac{1}{4}M & M \text{ even} \\ \frac{M^2}{8} - \frac{1}{8} & M \text{ odd} \end{cases}, \quad \left|F^{\sigma^l,M}_{Q,Q}\right| = \begin{cases} \frac{M^2}{8} + \frac{1}{4}M & M \text{ even} \\ \frac{M^2}{8} - \frac{1}{8} & M \text{ odd} \end{cases}. \tag{43}$$

The proof of orthogonality is done in a similar manner as in [1] and [3]. The result

$$\langle \varphi_\lambda, \varphi_\nu \rangle^{\sigma,M}_{Q,P^\vee} = 8M^2 h^M_Q(\lambda) \delta_{\lambda,\nu}, \tag{44}$$

for any $\lambda, \nu \in \Lambda^{\sigma,M}_{Q,P^\vee}$ proves that the orbit functions indexed by elements of $\Lambda^{\sigma,M}_{Q,P^\vee}$ constitute an orthogonal basis of the vector space of complex functions on $F^{\sigma,M}_{Q,P^\vee}$. The point grid for $M = 5$ is shown in Figure 3.

6.3 The cases $(P, P^\vee)$ and $(P^\vee, P)$

The case $\Lambda^{\sigma,M}_{P,P^\vee}$ has already been studied in [4] and [3]. The resulting orthogonality relations are

$$\langle \varphi_\lambda, \varphi_\nu \rangle^{\sigma,M}_{P,P^\vee} = 16M^2 h^M_Q(\lambda) \delta_{\lambda,\nu}, \tag{45}$$

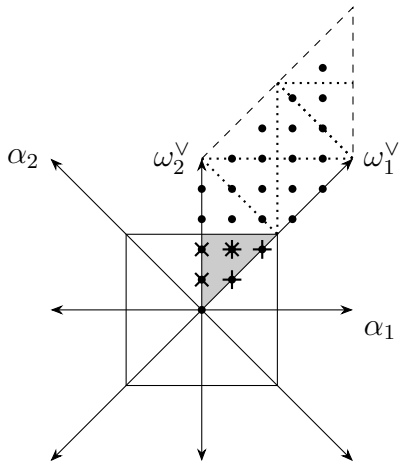


Figure 3: The point grids in the case $(Q, P^\vee)$ for $M = 5$. The grey triangle is the fundamental domain of the group $P^\vee \rtimes W$. The 25 black dots are the representants of the classes in $\frac{1}{5}P^\vee/P^\vee$. All the points lying in the grey triangle are distinct elements of $F^{\mathbb{1},5}_{P^\vee,P^\vee}$. The singular point in the interior of the grey triangle is the one element of $F^{\sigma^e,5}_{P^\vee,P^\vee}$. The points crossed with + are the elements of $F^{\sigma^s,5}_{P^\vee,P^\vee}$ and the points crossed with $\times$ are the elements of $F^{\sigma^l,5}_{P^\vee,P^\vee}$. The borders of the set $S_{P^\vee}$ are the lines ending with the arrows of the vectors $\omega_1^\vee$ and $\omega_2^\vee$ along with the opposing dashed lines; each dashed line is equivalent with the opposing non-dashed side. The triangles meeting in the origin are obtained by the action of W on the fundamental domain $F_{P^\vee}$. The dotted lines denote the borders of these triangles shifted to $S_{P^\vee}$, i.e. the images of $F_{P^\vee}$ under the action of W on $\mathbb{R}^2/P^\vee$. An analogous image of the label grid can be obtained by rescaling the gray triangle by the factor 5.

for $\lambda, \nu \in \Lambda^{\sigma,M}_{P,P^\vee}$. It holds that $\left|F^{\sigma,M}_{P,P^\vee}\right| = \left|\Lambda^{\sigma,M}_{P,P^\vee}\right|$, so the orbit functions labelled by elements of $\Lambda^{\sigma,M}_{P,P^\vee}$ form an orthogonal basis of the space of complex functions on $F^{\sigma,M}_{P,P^\vee}$.

This time, the dual case $(P, P^\vee)$ cannot be obtained from the previous one by simple multiplication, still we have

$$\left|F^{\sigma,M}_{P,P^\vee}\right| = \left|\Lambda^{\sigma,M}_{P^\vee,P}\right| \quad \text{and} \quad \left|\Lambda^{\sigma,M}_{P,P^\vee}\right| = \left|F^{\sigma,M}_{P^\vee,P}\right|, \tag{46}$$

so the cardinalities of the label set and point set are equal.

One can arrive to a completely analogous orthogonality relation

$$\langle \varphi_\lambda, \varphi_\nu \rangle^{\sigma,M}_{P^\vee,P} = 16M^2 h^M_{Q^\vee}(\lambda)\delta_{\lambda,\nu} \tag{47}$$

for any $\lambda, \nu \in \Lambda^{\sigma,M}_{P^\vee,P}$ by rescaling and switching the roles of the grids and following the procedure outlined in [4] and [3].

6.4 The case (P, P)

Note that this case is self-dual in the sense that the labelling and point lattices are equal. Clearly, it holds that $\left|F^{\sigma,M}_{P,P}\right| = \left|\Lambda^{\sigma,M}_{P,P}\right|$ and the orthogonality relation

$$\langle \varphi_\lambda, \varphi_\nu \rangle^{\sigma,M}_{P,P} = 32M^2 h^M_{Q^\vee}(\lambda)\delta_{\lambda,\nu} \tag{48}$$

holds for any $\lambda, \nu \in \Lambda^{\sigma,M}_{P,P}$, as was shown in [5].

7 Conclusion

We have given an overview of all the possible constructions of orthogonal sets of orbit functions in the root system C_2 using the lattices $P, Q, P^\vee$ and $Q^\vee$. Three cases out of the total six have already been examined and the orthogonality of orbit functions is known.

Out of the remaining three cases in C_2, $\Lambda^{\sigma,M}_{Q^\vee,P}$ degenerates into a simple rescaling of its dual case which has already been studied in [1]. Another case which has not been studied is $\Lambda^{\sigma,M}_{P^\vee,P}$ which is the dual of an already examined case [3, 4]. The orthogonality relation can be found using a completely analogous procedure. The final unstudied case is $\Lambda^{\sigma,M}_{Q,P^\vee}$, in which we were able to show the equal cardinality of the label and point sets and consequently find a discrete orthogonality relation for the orbit functions.

Given every possible combination of the label and point lattice, we have arrived at the orthogonality relation of the form

$$\langle \varphi^\sigma_\lambda, \varphi^\sigma_\nu \rangle^{\sigma,M}_{B,A} = 8M^2 h^M_{A^\perp}(\lambda) \left|A/B^\perp\right| \delta_{\lambda,\nu}. \tag{49}$$

The focus of our research at the moment is to prove that a relation of this type holds for any irreducible crystallographic root system along with equal cardinalities of the point and label sets. These orthogonality relations can be used as building blocks for Fourier-type transformations of functions on various n-dimensional grids which could be interesting from a computational perspective. Moreover, the theory of orbit functions has application in physics, for example, in modelling of the properties of graphene [2].

References

[1] J. Hrivnák and L. Motlochová. *Dual-root lattice discretization of Weyl orbit functions.* Journal of Fourier Analysis and Applications **25** (2019), 2521–2569.

[2] J. Hrivnák and L. Motlochová. *On electron propagation in triangular graphene quantum dots.* Journal of Physics A: Mathematical and Theoretical **55** (2022), 125201.

[3] J. Hrivnák, L. Motlochová, and J. Patera. *On discretization of tori of compact simple Lie groups: II.* Journal of Physics A: Mathematical and Theoretical **45** (2012), 255201.

[4] J. Hrivnák and J. Patera. *On discretization of tori of compact simple Lie groups.* Journal of Physics A: Mathematical and Theoretical **42** (2009), 385208.

[5] J. Hrivnák and M. A. Walton. *Weight-lattice discretization of Weyl-orbit functions.* Journal of Mathematical Physics **57** (2016).

[6] R. M. Kane. *Reflection groups and invariant theory*, volume 5. Springer, (2001).

[7] A. Klimyk and J. Patera. *Orbit functions.* SIGMA. Symmetry, Integrability and Geometry: Methods and Applications **2** (2006), 006.

[8] A. Klimyk and J. Patera. *Antisymmetric orbit functions.* SIGMA. Symmetry, Integrability and Geometry: Methods and Applications **3** (2007), 023.

[9] R. V. Moody, L. Motlochova, and J. Patera. *Gaussian cubature arising from hybrid characters of simple Lie groups.* Journal of Fourier Analysis and Applications **20** (2014), 1257–1290.

Power Spectral Features of Musculoskeletal Disorders and Their Classification*

Nichita Vatamaniuc
vatamnic@fjfi.cvut.cz

study programme: Applied Informatics
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. This paper focuses on processing the data gained from human body motion sensors followed by the recognition of patients who are suffering from a neurological disease that leads to musculoskeletal system disorder. The frequency spectrum acquired from the processed signal generated by human movement is then used to get the power spectral features for each patient. These features are then used for binary classification to recognize if the patient is healthy or has a locomotion problem.

Keywords: frequency spectrum, classifier, frequency features, biomedical diagnosis, musculoskeletal disorder.

Abstrakt. Tento článek je zaměřen na zpracování dat získaných ze senzoru pochybu umístěných na lidském těle. Zároveň lékař určil diagnózu neurologického onemocnění, které vede k poruchám pohybového aparátu. Ze zpracovaného signálu bylo získáno frekvenční spektrum pochybu člověka a využito ke konstrukci spektrálních charakteristik každého pacienta. Tyto charakteristiky byly použité pro klasifikaci do dvou tříd a k rozpoznání zda pacient je zdrav nebo má problémy z pohybem.

Klíčová slova: Frekvenční spektrum, klasifikátor, frekvenční charakteristiky, biomedicínská diagnóza, nemoci pohybového aparátu.

1 Introduction

1.1 Neurological diseases and Musculoskeletal disorders

The musculoskeletal system also known as the locomotor system is one of the human body systems which grants movement and stability to the human body. The elements of the system participate in nervous regulation of locomotion, facial expressions, maintenance of posture, and other processes and functions such as cushioning shocks and concussion, and protecting vital organs or metabolism. [11]. Unfortunately, there are diseases and disorders that harmfully act on the musculoskeletal system. Usually, these diseases are difficult to diagnose due to the fact that the musculoskeletal system is in close relation with the other human body systems [16]. In this paper, the focus was

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS23/190/OHK4/3T/14.

shifted to musculoskeletal disorders which were caused by neurological disorders, since they are common companions for locomotion problems. Neurological disorders result in structural, biochemical, or electrical abnormalities in the human nervous system. Usually, problems arise in the brain or spinal cord with typical symptoms such as paralysis, muscle weakness, and poor coordination [8]. An example of such a disease that affects the motor system is Parkinson's disease. Symptoms of locomotion problems occur due to the dying of nerve cells in the region of midbrain [2].

1.2 Diagnosis and experiment aim

The most common way of diagnostics of musculoskeletal and neurological disorders is the analysis of balance problems by testing the gait since these disorders proved to be a ground for an abnormal gait [1]. The problems with human locomotion stability rise dramatically with age. It can be challenging to detect abnormalities in gait by reason of the fact that there are no accepted standards for detecting abnormalities amongst young and old generations [12]. A further problem, regardless of the existing of diagnosis method problem, is the manual analysis of the data. The core of this problem is based not only on being time-consuming and expensive but significantly on being dependent on the doctor's experience as well. However, different research shows that machine learning techniques can be effective in the automatization of gait analysis for diagnosing musculoskeletal and neurological disorders [17]. In contrast to other approaches, this paper is targeted at classifying the power spectrum features extracted from the frequency spectrum. The frequency spectrum is the result of the signal processing captured from the motion sensor applied to the patient. Furthermore, the data used for the experiment contains not only a gait test. The patients were asked to perform gait and stance tests that are included in a scale for the assessment and rating of ataxia (SARA) [10] with the motion sensor. The gait test required the patient to walk parallel to a wall and then turn around in the opposite direction of gait and to walk in tandem (heels to toes) without support. The stance test required the patient to stand in a natural position with feet together in parallel, after in tandem with open eyes. These test clearly defines the ability of the patient to keep the body balance. The dataset contains signals captured from 51 patients (Table 1).

The signal from the sensor consists of nine channels. These channels represent the relative change of acceleration, gyroscope, and angle in a three-dimensional space. The sensor is set to capture a signal with a sampling period of 0.02 s. The aim of this experiment is to develop a novel method of recognition of patients with possible musculoskeletal disorders caused by neurological diseases with the help of the power spectrum features extracted from the frequency spectrum of the motion sensor signal.

2 Signal processing and feature extraction

2.1 Noise reduction and segmentation

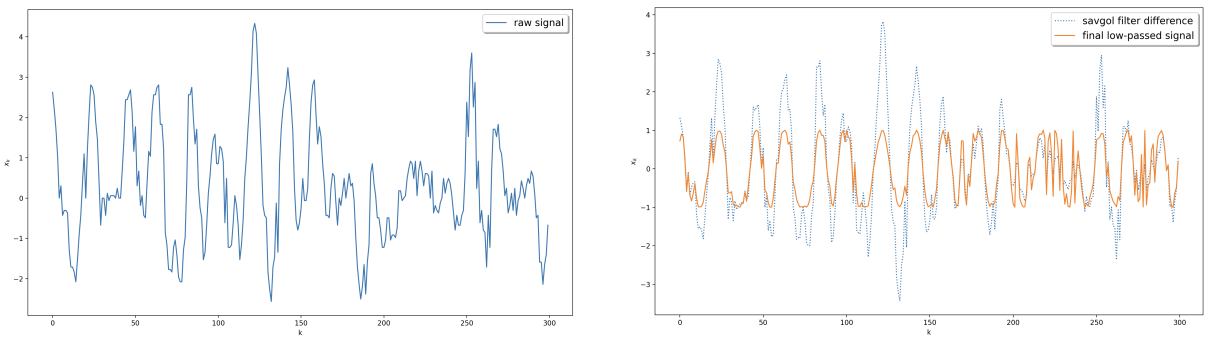
It is an inefficient approach to work with the raw signal from the sensor. In most cases, the raw signal is very noisy and has a lot of useless information (Figure 1a) which will

Table 1: Data cardinality.

Disease	Patients	Gait tests	Stance tests
Vestibular Sindrome	1	3	4
Tumor	1	3	4
Stroke	13	35	47
Spastic Paraparesis	1	0	4
Parkinsonism	2	3	4
NPH	1	2	3
Myelopathy	2	1	3
MS	2	6	8
Healthy	16	42	60
Cerebellar Dysfunction	2	4	6
ALS	1	1	4
Polyneuropathy	9	15	19

increase the computational cost and decrease the final accuracy of classification methods. Another important step is to define which features will be extracted from the signal and in what form they will be represented. Accordingly, it is very important to reprocess the raw signal to get significantly better results.

As described in the previous section, the sampling period of the sensor is $T_\mathrm{s} = 0.02$ s, which gives us the signal frequency $f_\mathrm{s} = \frac{1}{T_\mathrm{s}} = 50$ Hz. Let us assume that the number of samples in one record is $N \in \mathbb{N}$. Thus, the k-th element is $x_k \in \mathbb{R}$. The sample is defined as $\{x_k\}_{k=0}^{N-1}$. In order to make the signal sample smoother the Savitzky-Golay filter [6] is applied. Hereafter, the filtered result is subtracted from the original sample. The frequency response of the smoothed signal is then cut off with the low-pass filter, which passes the signal with a frequency lower than a selected cutoff frequency and attenuates signals with frequencies higher than the cutoff frequency. This was achieved by generating the Hilbert envelope [18] of the smoothed signal and dividing the smoothed signal by envelope. The final step is to get filtered and normalized data to $[-1, 1]$. Solid line on the Figure 1b.



(a) Raw signal.

(b) Filtered signal.

Figure 1: Signal preprocessing steps. Signal cutoff to 300 samples.

Smoothed and filtered signals now can be divided into segments. This step is required for getting power spectrum features in further steps. The signal will be segmented into the window system, where $w \in \mathbb{N}$ is a window length. Then the k-th element of the j-th segment is defined as $\{x_{k+jw}\}_{k=0}^{w-1}$. The total number of segments is $M = \lfloor \frac{N}{w} \rfloor \in \mathbb{N}$ and the remaining data are neglected.

2.2 Segment preprocessing

Any segment is presented as a fixed length vector $(\xi_0, \ldots, \xi_{w-1})$. The goal to achieve in this part is to calculate the power spectrum of the given segment. The power spectrum of a segment is $(P_0, \ldots, P_{w-1})$ where $P_k = |\Psi_k|^2$ and Ψ_k is k-th component of Fourier image obtained by Discrete Fourier Transform (DFT) [9]

$$\Psi_k = \sum_{n=0}^{w-1} \xi_n \cdot \exp\left(\frac{-\mathrm{j}2\pi kn}{w}\right)$$

For better time and calculation performance we can use $w = 2^m, m \in \mathbb{N}$. Thereby, the Fast Fourier Transform (FFT) [9] can be used instead. When previous calculations were done, the j-th segment is presented as $\vec{d_j} = (P_1, \ldots, P_{\frac{w}{2}-1})$, since the magnitude of the Fourier Transform is symmetric, only $\frac{w}{2} - 1$ elements are taken (Figure 2a). In this study, we use segment length $w = 100$.

2.3 Final feature reprepsentation

When all segments are processed, the power spectrum of a signal sample is

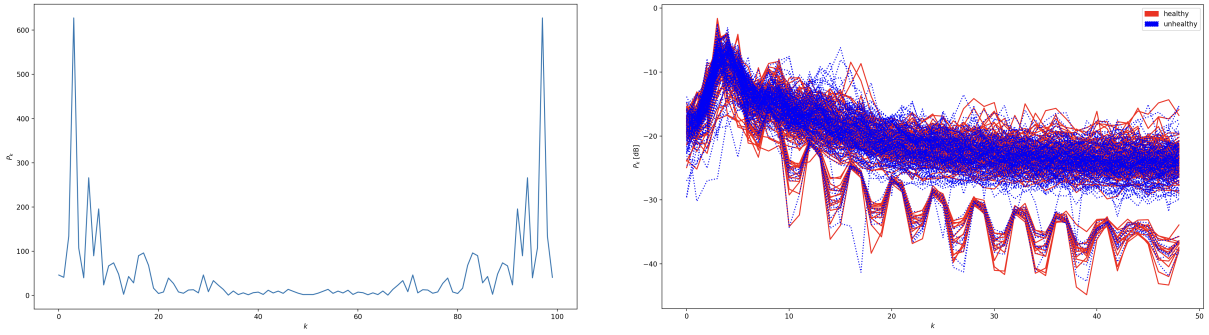
$$\vec{h} = \frac{1}{M} \sum_{j=1}^{M} \vec{d_j},$$

which is a component-wise arithmetic mean of segment powers (Figure 2b). A group of patients was investigated using various tests (gait, stance) and the pattern is a result of a single test on a given patient. Therefore, one patient can generate several patterns per test type. The sensor generates nine time series but we use only three channels from the gyroscope. The final pattern consists of three powers spectra ($c = 3$) as vector $(\vec{h_x}, \vec{h_y}, \vec{h_z})$.

3 Classification methods

3.1 Dimensionality reduction

For the experiment, the first four elements of each spectrum were cut off, since they do not contain any useful information as visualized in the Figure 2b. Additionally, the spectrum with corrupted data, which is at the bottom of the same Figure, was removed as well. The power spectrum feature of the patient which was used for further classification in this experiment consists of contamination of the power spectrum from three channels which are the three dimensions of the accelerometer. Since the segment length is $\frac{w}{2} - 1$ and the first four elements were cut, the patient feature is the vector of length $L =$



(a) FFT application for $w = 100$.

(b) Vizualization of all segments for $w = 100$.

Figure 2: Segment processing.

$((\frac{w}{2} - 1) - 4) \cdot c = 45 \times 3 = 135$. In order to reduce the dimensionality of the problem Principal Components Analysis [14] were used. Each patient feature vector size was reduced from 135 to 39 since this number of features gave the best result after multiple attempts with different parameters during classification. Figure 3 shows the explained variance ratio of each component.

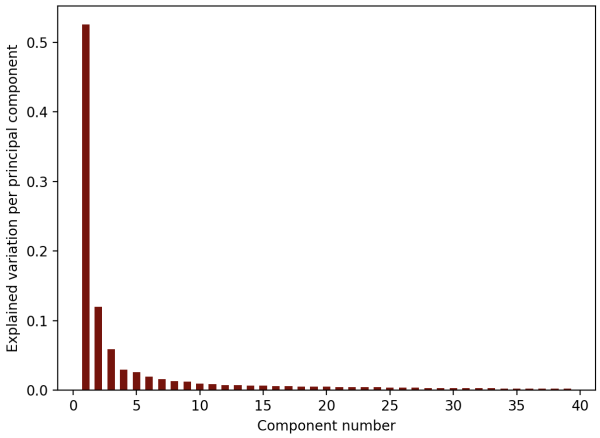


Figure 3: Explained variance ratio.

3.2 Multilayer Perceptron

The first method for the classification problem in this experiment was a multilayer perceptron classifier. The approach is based on supervised learning of an artificial neural network (ANN). The training dataset is processed by ANN to build a function that maps new data on expected output values. Accordingly, each patient's feature was labeled with 0 - unhealthy and 1 - health. Thus, during the training on the dataset, it will be possible to distinguish which category belongs to the patient. The structure of the ANN used for the purposes of this experiment is the following:

- Input layer of size 39
- Four hidden layers of size (32,32,16,8)

- Output binary layer with a simple neuron

The ANN model is trained through the backpropagation. It performs a backward pass to adjust the ANN's weights and biases to minimize mean squared error [3]. Adaptive Moment Estimation (Adam) was taken as the optimization algorithm for the backpropagation. The advantage of the Adam algorithm is that it computes adaptive learning rates for each parameter in contrast to stochastic gradient descent which does not change the learning rate during training [4]. Another important feature of the ANN is the activation function which is responsible for calculating the sum of the product of weights and inputs with bias defining the neuron output in a range of values. For this experiment, the Rectifier Linear Unit (ReLU) was selected. The main advantage of the ReLU is that it is capable of resolving the vanishing gradients problem [19]. Corresponding formula is

$$\text{relu}(x) = \max(0, x)$$

3.3 Support Vector Machine

Support Vector Machine (SVM) is another classification method that was used in this experiment for the classification task. The main principle of the SVM is to find a hyperplane that will be maximally distant from the data points from different clusters and at the same time will separate data into different clusters [15]. For high-dimensional space problems, SVM can be extended with a kernel function. The kernel function is a mathematical function that maps data from one feature space into another, thus allowing linear classifiers to deal with nonlinear tasks. For the experiment, the choice was made in favor of the radial basis function kernel (RFB) [7]. The formula of this kernel is

$$K(x, x') = \exp(-\gamma||x - x'||^2)$$

where γ is a parameter for similarity measuring between two points and equals to $\frac{1}{2\sigma^2}$, with σ as a free parameter and $||x - x'||^2$ as squared Euclidean distance between two feature vectors. The best result was achieved with $\gamma = 0.841$.

4 Results

The whole experiment was made with the Python programming language. For the data preprocessing was used SciPy library [13] and for Classification methods Scikit-learn library [5]. Trained classifiers were tested with cross-validation. Available data was separated into two sets. One set was used for training and another for testing. Since there was lack of data to train, 85% of data was applied for training and the remaining 15% for testing. The accuracy of the ANN classifier is 0.789, with the following confusion matrix

$$\begin{bmatrix} 22 & 3 \\ 5 & 8 \end{bmatrix}$$

with sensitivity to unhealthy $\frac{22}{27} = 0.814$ and to healthy $\frac{8}{11} = 0.727$. Need to mention, that the RBF kernel in the Scikit-learn library has another parameter that changes the rate between wrong classification and decision hyperplane simplicity and it was set to $c = 1.299$. The SVM classifier ended up with a better result. The accuracy is 0.842. The confusion matrix looks as follows

$$\begin{bmatrix} 24 & 1 \\ 5 & 8 \end{bmatrix}$$

from which we have sensitivities $\frac{24}{29} = 0.827$ and $\frac{8}{9} = 0.888$ for unhealthy and health accordingly. The SVM classifier resulted better in the classification of unhealthy patients.

5 Conclusion

This paper reveals the fact that power spectrum features extracted from the motion sensor signal can be used for diagnosing a neurological disease that affects the musculoskeletal system. Such neurological diseases can affect human body balance and cause locomotion problems that are distinctive from healthy human body signals. The difference between healthy and unhealthy patients can be considered as an abnormality. As discovered during the experiment, such abnormalities can be classified by different classification algorithms. Two different methods for classification were tested in this study. The SVM classifier had a slightly better result than the ANN classifier. However, there are a lot of different ways to configure the ANN classifier which can improve the final result. On the other hand, signal preprocessing parameters like window length or number of components can be configured as well and that can lead to overall better results in classification for any method. Nevertheless, the main goal of the experiment was to show that power spectrum features can be used for detecting abnormalities in human locomotion, and it was achieved.

References

[1] B.-C. Lee, A. Fung and T. Thrasher, *The effects of coding schemes on vibrotactile biofeedback for dynamic balance training in Parkinson's disease and healthy elderly individuals*, IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 1, pp. 153-160, (2018).

[2] B. R. Bloem, M. S. Okun, and C. Klein, *Parkinson's disease.* The Lancet, 397(10291), 2284-2303. (2021).

[3] D. E. Rumelhart, R. Durbin, R. Golden, and Y. Chauvin. *Backpropagation: The basic theory.* In Backpropagation, pp. 1-34. Psychology Press, (2013).

[4] D.P. Kingma, and J. Ba. *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980 (2014).

[5] F. Pedregosa, V. Gaël, A. Gramfort, V. Michel, T. Bertrand, O. Grisel, M. Blondel et al. *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research 12: 2825-2830. (2011).

[6] H. William, and A. Saul Teukolsky. *Savitzky-Golay smoothing filters.* Computers in Physics 4, no. 6: 669-672. (1990).

[7] K. Thurnhofer-Hemsi, L. Ezequiel, M. A. Molina-Cabello, and N. Kayvan. *Radial basis function kernel optimization for support vector machine classifiers.* arXiv preprint arXiv:2007.08233 (2020).

[8] K. T. Thakur, E. Albanese, P. Giannakopoulos, N. Jette, M. Linde, M. J. Prince, ... and T. Dua, *Neurological disorders. Disease Control Priorities*, 4, 87-107. (2016).

[9] J. Nussbaumer, Henri, and J. Henri Nussbaumer. *The fast Fourier transform.* Springer Berlin Heidelberg, (1982).

[10] M. Grobe-Einsler, A.T. Amin, J. Faber, et al. *Scale for the Assessment and Rating of Ataxia (SARA): Development of a Training Tool and Certification Program.* Cerebellum (2023).

[11] M. Nordin, and V. H. Frankel, *Basic biomechanics of the musculoskeletal system.* Lippincott Williams & Wilkins. (2001).

[12] N. B. Alexander, *Gait disorders in older adults*, J. Amer. Geriatrics Soc., vol. 44, no. 4, pp. 434-451, (1996).

[13] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski et al. *SciPy 1.0: fundamental algorithms for scientific computing in Python.* Nature methods 17, no. 3: 261-272. (2020).

[14] R. Bro, and K. Age, Smilde, *Principal component analysis.* Analytical methods 6, no. 9: 2812-2831. (2014).

[15] S. Suthaharan, and S. Suthaharan. *Support vector machine.* Machine learning models and algorithms for big data classification: thinking with examples for effective learning: 207-235. (2016).

[16] S. W. Driscoll, and J. Skinner, *Musculoskeletal complications of neuromuscular disease in children.* Physical Medicine and Rehabilitation Clinics of North America, 19(1), 163-194. (2008).

[17] T. D. Pham, *Texture classification and visualization of time series of gait dynamics in patients with neuro-degenerative diseases*, IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 1, pp. 188-196, (2018).

[18] U. Timothy, *Envelope calculation from the Hilbert transform.* Los Alamos Nat. Lab., Los Alamos, NM, USA, Tech. Rep (2006).

[19] X. Glorot, A. Bordes, Y. Bengio, *Deep sparse rectifier neural networks.* AISTATS. (2011).

Burgers'-Type Equation As a Model of Reaction-Diffusion Pattern Formation

Hazal Yurtbak
hazal.yurtbak@fjfi.cvut.cz

study programme: Mathematical Engineering
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. The study of reaction-diffusion systems has been crucial in various scientific disciplines due to their ability to explain natural occurrences like pattern formation, wave propagation, and oscillations. The Burger's type equation, a nonlinear partial differential equation that combines diffusion and advection terms, has become a prominent tool for modeling reaction-diffusion pattern formation. Its strength lies in its capability to illustrate complex patterns, making it applicable to diverse fields like population dynamics, fluid dynamics, and chemical reactions. Additionally, it has enhanced our understanding of pattern formation in the context of non-equilibrium thermodynamics. In this paper, we explore the Burger's type equation's role in modeling reaction-diffusion patterns, discussing its fundamental principles, connections to Turing models and reaction-diffusion systems, and its contributions to pattern formation studies. We will also delve into numerical methods for solving these equations and obtaining solutions.

Keywords: burger's equation, pattern formation, reaction-diffusion, turning model

1 Introduction

Reaction-diffusion systems have been an important area of research in many scientific fields including mathematics, physics, chemistry, and biology. The study of these systems has been largely motivated by their ability to describe a variety of natural phenomena, such as pattern formation, wave propagation, and oscillations. One of the most widely studied models of pattern formation is the Turing model, which is based on a set of partial differential equations that describe the dynamics of interacting chemicals. However, this model has limitations in its ability to capture complex patterns that are often observed in natural systems. In recent years, the Burger's type equation has emerged as a promising model for studying reaction-diffusion pattern formation. This equation is a nonlinear partial differential equation that incorporates both diffusion and advection terms, which allows for the description of a wider range of phenomena than the Turing model. The Burger's type equation has been used to study a variety of systems, including population dynamics, fluid dynamics, and chemical reactions. One of the main advantages of the Burger's type equation is its ability to describe the formation of complex patterns that are not captured by simpler models. In particular, it has been shown to be capable of describing the formation of traveling waves, spatiotemporal chaos, and other patterns that

arise in many natural systems. This has led to significant interest in the use of the Burger's type equation as a tool for understanding pattern formation in a variety of scientific fields. The Burger's type equation has also been used in conjunction with the principles of non-equilibrium thermodynamics, which provides a framework for understanding the behavior of systems that are far from equilibrium. This has led to new insights into the relationship between pattern formation and the principles of non-equilibrium thermodynamics.

In this paper, we will explore the use of the Burger-type equation as a model for reaction-diffusion pattern formation. Focusing on the Brusselator model, we will examine the fundamental principles of this equation, its relationships with Turing models, and its utilization in pattern formation studies. Furthermore, we will review the connection between this equation and the principles of non-equilibrium thermodynamics, offering a novel perspective on its relationship with pattern formation. This study aims to contribute to a better understanding of pattern formation across various scientific disciplines.

References

[1] V. Klika. *Pattern formation revisited within non-equilibrium thermodynamics:Burgers'type equation.* Biological Cybernetics (2022), 116:81–91.

[2] A. M. Turing. *The chemical basis of morphogenesis.* The Royal Society **237** (1952), no 641.

[3] I. Prigogine - G. Nicolis. *Self-Organisation in Nonequilibrium Systems: Towards A Dynamics of Complexity.* Bifurcation Analysis (1985), pp.3-12.

Non-equilibrium Strain and Elastic Hysteresis in Static and Dynamic Experiments in Sandstones

Radovan Zeman
zemanra5@fjfi.cvut.cz

study programme: Mathematical Engineering
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 of the Czech Academy of Sciences

Jan Kober, Department of Impact and Waves in Solids
Institute of Thermomechanics of the Czech Academy of Sciences

Abstract. Consolidated granular materials exhibit intriguing characteristics in their elastic response, including slow dynamic behavior and hysteresis. This hysteresis, which manifests as distinct elastic properties during loading and unloading, is evident in both quasi-static experiments, where slow loading and unloading cycles are employed, and dynamic acousto-elastic testing, where rapid perturbations are induced using propagating or standing waves.

The underlying cause of this hysteresis in the elasticity of consolidated granular materials continues to be a subject of active debate. In our study, we attribute this hysteresis to the influence of slow dynamics, which simultaneously introduces elastic anisotropy due to nonlinear effects, particularly affecting waves polarized and propagating parallel to the applied load. We explain this slow dynamics through the concept of non-equilibrium strain, which gradually accumulates within the material under the influence of induced strain and slowly dissipates as the applied strain diminishes.

The experimental investigations, conducted on sandstone samples, span an extensive range of strain levels, encompassing five orders of magnitude. This range extends from the dynamic regime, where strain levels are on the order of 10^{-7} due to excitation of the first compressional mode, to strain levels of 10^{-2} induced by uniaxial compression using a tensile testing machine. To monitor variations in velocity, we employ high-frequency longitudinal and shear wave transducers, utilizing pulses with various polarizations propagating in the transverse direction. Our findings provide compelling evidence that the proposed model effectively accounts for the observed behavior across this entire strain spectrum.

Keywords: acoustoelastic testing, consolidated granular materials, elastic hysteresis, nonlinear elasticity

Abstrakt. Konsolidované granulované materiály vykazují charakteristické elastické chování včetně slow dynamics a hystereze. Tato hystereze, která se projevuje odlišnými elastickými parametry při zatěžování a odlehčování, je patrná jak při kvazistatických experimentech, kdy se provádějí pomalé cykly zatěžování a odlehčování, tak při dynamickém akusticko-elastickém testování, kdy se rychlé perturbace vyvolávají pomocí šířících se nebo stojatých vln.

Příčina hystereze v elasticitě konsolidovaných granulovaných materiálů je stále předmětem diskusí. V naší práci přisuzujeme tuto hysterezi vlivu slow dynamics, který současně vyvolává elastickou anizotropii v důsledku nelineárních efektů, zejména ovlivňujících vlny polarizované a šířící se rovnoběžně s působícím zatížením. Slow dynamics vysvětlujeme pomocí konceptu nerovnovážné deformace, která se postupně hromadí v materiálu pod vlivem indukované deformace a pomalu se vytrácí po té, co se aplikovaná deformace zmenšuje.

Experimenty provedené na pískovcových vzorcích pokrývají rozsah pěti řádů deformací, od dynamického zatěžování, kde se úrovně deformace pohybují v řádu 10^{-7} v důsledku excitace prvního podélného módu, po deformace 10^{-2} vyvolané jednoosým stlačením pomocí systému pro tahové zkoušky. Ke sledování změn rychlosti jsou použity vysokofrekvenční snímače podélných a smykových vln, které vysílají pulzy s různou polarizací šířící se v příčném směru. Výsledky potvrzují, že navržený model vysvětluje pozorované chování v celém spektru deformace.

Klíčová slova: akustoelastické testování, elastická hystereze, konsolidované granulované materiály, nelineární elasticita

Full paper: R. Zeman, J. Kober, M. Scalerandi, *Non-equilibrium Strain and Elastic Hysteresis in Static and Dynamic experiments in sandstones.* To be published in the Conference Proceedings of Forum Acusticum 2023.

Optimization of Fast Parallel Operations with Large Disk Arrays*

Martin Zemko
martin.zemko@fjfi.cvut.cz

study programme: Applied Informatics
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. This paper addresses the critical need for reliable and efficient data storage in the context of a high-energy physics experiment called AMBER. The experiment generates massive data rates of up to 10 GB/s, requiring the optimization of data archiving and retrieval systems. The study investigates single-disk performance, including random and sequential reads, highlighting the impact of parallel access and disk geometry. A comparison with solid-state drives (SSD) reveals important differences. The paper then introduces the concept of redundant arrays of independent disks (RAID) and assesses various RAID configurations, considering factors such as reliability, data rates, and capacity. Probability analysis is used to evaluate the success of RAID rebuilding in the event of disk failure. In addition, an innovative approach of alternating disk access is proposed to ensure uninterrupted performance in case of disk failures. Finally, the study identifies the most suitable RAID configuration for the high energy physics experiment requirements. The results contribute to the design of high-performance storage solutions for data-intensive scientific experiments, balancing performance, redundancy, and capacity.

Keywords: data storage, RAID configurations, parallel data access, disk geometry

Abstrakt. Tento článek se zabývá potřebou spolehlivého a efektivního ukládání dat v kontextu experimentu fyziky vysokých energií nazývaného AMBER. Experiment generuje obrovské množtsví dat dosahující až 10 GB/s, což vyžaduje optimalizaci systémů pro archivaci a odečítání dat. Tato studie zkoumá výkon jednoho disku, včetně náhodného a sekvenčního čtení, a zdůrazňuje vliv paralelního přístupu a geometrie disku. Srovnání se solid-state disky (SSD) odhaluje důležité rozdíly. Článek se dále zabývá konceptem redundantních polí nezávislých disků (RAID) a hodnotí jejich různé konfigurace zohledňující faktory jako spolehlivost, přenosovou rychlost a kapacitu. Pravděpodobnostní analýza je použita k posouzení úspěšnosti obnovy pole RAID v případě selhání disku. Kromě toho je navržen inovativní přístup střídavého přístupu k diskům, aby byl zajištěn nepřetržitý provoz v případě selhání disku. Nakonec studie identifikuje nejvhodnější konfiguraci RAID pro potřeby experimentu v oblasti vysokoenergetické fyziky. Tyto zjištění přispívají k návrhu výkonných úložných řešení pro datově náročné vědecké experimenty, která zohledňují výkon, redundanci i kapacitu.

Klíčová slova: ukládání dat, RAID konfigurace, paralelní přístup k datům, geometrie disků

*This work has been supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS23/190/OHK4/3T/14.

1 Introduction

Reliable storing and processing of large data quantities is an integral part of every experiment in high-energy physics. Particle detectors comprising millions of data channels produce excessive data rates and pose significant demands on computing infrastructure, including data storage. To accommodate the need for reliable and fast data archival and retrieval, complex systems of disk arrays are being used and exploited to their maximum capabilities. The architecture, organization, and configuration of these systems play an important role in reaching the best performance with available hardware.

The AMBER experiment located in the CERN laboratory is one such experiment using a triggerless readout system with online data filtering. According to initial estimations, the full-scale detector setup will produce the data rate reaching 10 GB/s of sustainable data flow[7]. Storing and accessing these quantities in real-time requires optimization on many levels – data organization on disks, access patterns, file system, OS-level optimizations, etc. The disk storage will decouple the online computing system that requires up to 100 % uptime from the data reduction system that can tolerate one or two days of downtime. The storage space will act as a temporary buffer, providing clear separation between both systems.

Describing our test setup, the storage system relies on 10 readout host servers, each equipped with an external storage chassis providing 24 disk bays. These chassis are connected to their host computers via Broadcom MegaRAID SAS 9380-8e RAID controller using fast 2 x 12Gb/s mini-SAS links providing sufficient throughput. The RAID controller supports various configurations that are analyzed and evaluated later in this paper. Installed drives are industry-grade spinning disks (MG07ACA14TE) manufactured by Toshiba. This model uses the world's first 9-disk design filled with helium. The helium-sealed environment reduces aerodynamic drag and lowers the power consumption. Each drive provides up to 14 TB of raw capacity using conventional magnetic recording. Their primary specification aims at cloud storage, business-critical servers, and object storage solutions. [6] In total, the fully-equipped single storage unit provides 336 TB of raw data capacity. The AMBER experiment will employ at least eight such storage systems providing more than 2 PB of disk capacity.

2 Single disk performance

A basic method of optimizing data processing tasks is to split data into smaller chunks and process them in parallel. The same method can be used to optimize the performance of storage systems. Let's assume a perfectly partitionable task that can be easily parallelized, and our goal is to measure only a single disk's performance by utilizing multiple workers to access the storage. Such a measurement gives us the characteristics of a single device.

2.1 Random reading

Spinning hard drives have a high seek latency that is caused by the physical positioning of the reading head moving to the requested location on the disk. Thus, HDD can serve

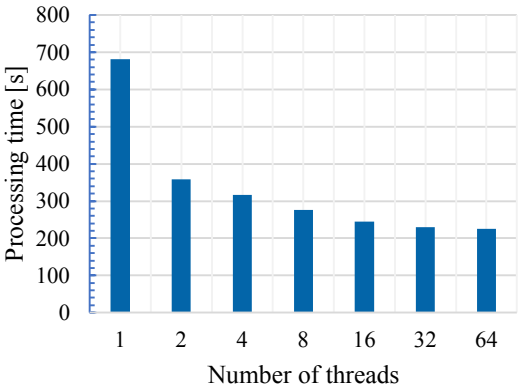


Figure 1: Random reading of 1 million files from HDD (lower is better)

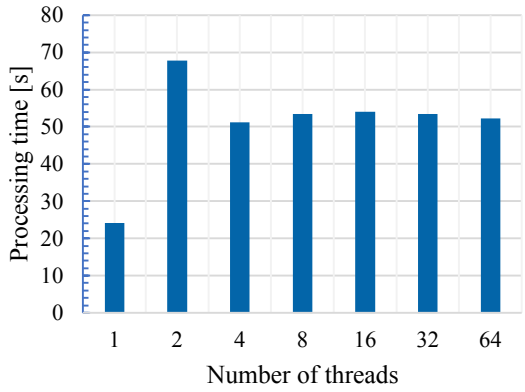


Figure 2: Sequential reading of 5 GB data file from HDD (lower is better)

only a single request at any given time. It can be assumed that the performance suffers from frequent changes in head position. However, according to our measurements shown in Figure 1, it is apparent that even for small random requests, there are significant gains in parallel access with multiple threads. This characteristic can be explained by optimizations performed by the operating system that can reorganize requests in the most efficient way, so the head minimizes its movement distance.

2.2 Sequential reading

On the other hand, when sequentially reading large data chunks, the situation is completely different. Parallel access significantly decreases the performance, and the total throughput is even lower than in single-threaded access. Figure 2 clearly illustrates that parallel access does not perform a proper optimization, and the performance is degraded. This is because the operating system interleaves IO requests coming from different threads, and the disk frequently repositions its head jumping from one position to another.

The most efficient way to solve this problem at the application level is to allow no more than one thread to access a single HDD device at any given time. Ideally, the application should emit its IO requests from a dedicated thread responsible for serializing the accesses. Other threads should use intra-process communication channels to acquire any resources from this dedicated thread. An alternative solution is based on mutexes that lock the resources only for one thread at a time. In this case, the granularity should be coarse enough to mitigate the impact of head seek time.

2.3 Comparison with solid-state drives

Unlike traditional HDDs, solid-state drives (SSD) can benefit from parallel processing regardless of sequential or random access. This is because their inner structure comprises several independent chips and data lanes serving multiple data streams. Our measurement (Figure 3) shows that parallel random reading has a significant impact on the overall

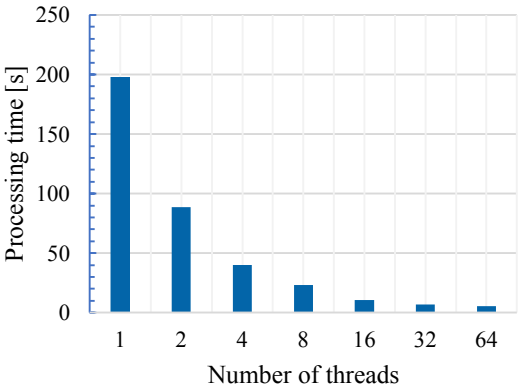


Figure 3: Random reading of 1 million files from SSD (lower is better)

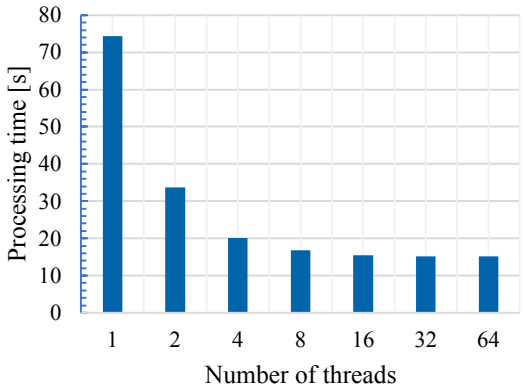


Figure 4: Sequential reading of 50 GB file from SSD (lower is better)

performance scaling almost linearly with the number of threads. However, there is a maximum throughput that limits the data rate for a higher number of threads (16 or more). For sequential access to SSD, benefits can be considerably high when reading large data files in a parallel manner. However, the plateau is observed earlier (see Figure 4).

2.4 Disk platter geometry

Spinning drives have another well-known issue caused by the geometry of rotating platters [4]. The main problem is that the performance decreases as the disk gets full. We observed such behavior in our past measurements, and we wanted to quantify the impact of this issue on the final performance. The measurement consisted of writing and reading at the maximum capacity of a single disk (14 TB). During this test, a relation between data rates and disk occupancy was measured.

Results show that reading or writing at the inner edge (fully occupied disk) decreases

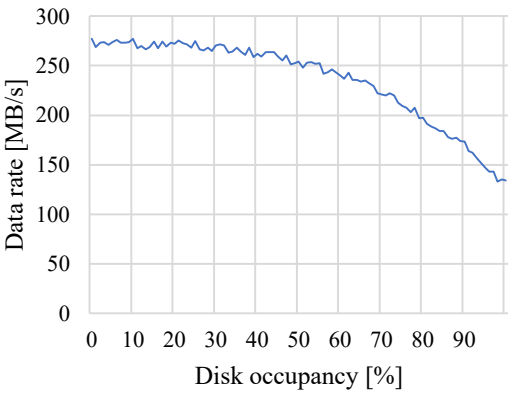


Figure 5: Sequential read from full HDD

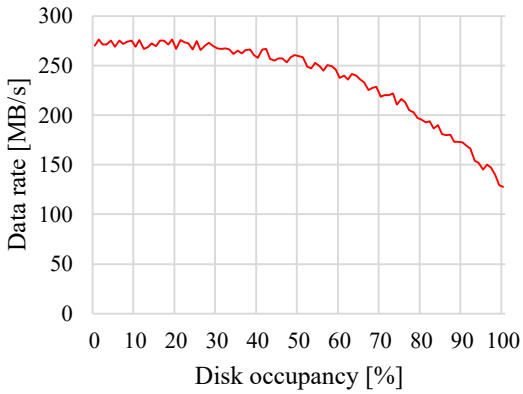


Figure 6: Sequential write from full HDD

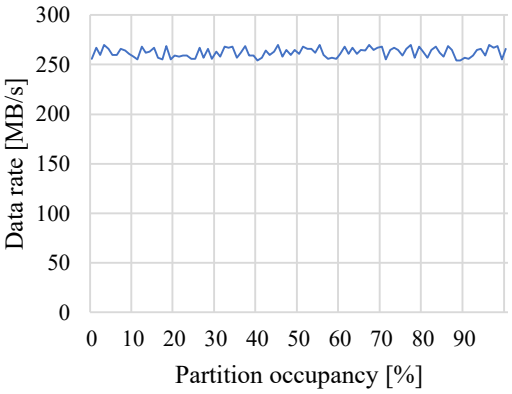


Figure 7: Sequential read from empty HDD

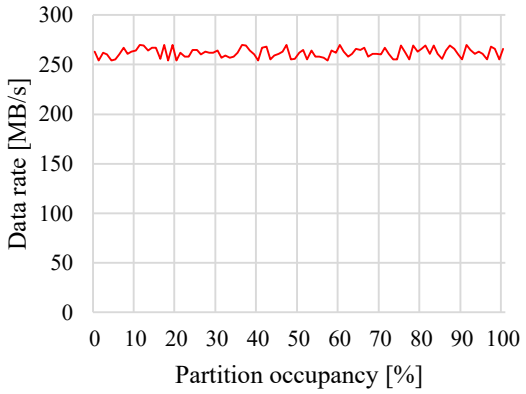


Figure 8: Sequential write from empty HDD

the throughput to approx. 48 % of the maximum data rate as observed on the outer region (empty disk). This issue is caused by the physical arrangement of data on the disk. Data are usually stored in concentric circles called cylinders. These cylinders get smaller as they approach the inner edge. Since the rotation speed is fixed (7200 rotations per minute) and the data density is constant across the entire disk, these facts result in a smaller capacity of inner cylinders and decreasing data rate towards the inner disk edge.

To verify our statement, we also created a small disk partition (1 TB) on the outer disk region and measured its performance. As can be seen in Figure 7 and Figure 8, the resulting data rate remains constant during the entire measurement, and the effects of disk geometry are negligible.

The outcome of these tests indicates that the performance of a single disk is not constant, and data rates depend on the head position above the platters. To achieve optimal results, one should create partitions and store data on the outer disk edge, where the best performance can be reached. Moreover, HDD disks clearly demonstrate their degradation as they get full. It is necessary to be aware of such quirks when designing a performance-critical storage solution, and one must add an additional size margin in order to avoid the full occupancy of disks.

3 Redundant array of independent disks

We measured the behavior of a single disk under different types of load. However, the throughput of a single disk does not scale linearly with the number of threads. To reach the desired data rate, it is necessary to coordinate access to multiple disks by multiple processes. A commonly used method is based on a virtualization technology called the redundant array of independent disks (or RAID in short). RAID is a data virtualization system combining multiple physical disks into a single logical unit called a volume. Thanks to this additional level of abstraction, volumes can provide extra features such as improved performance, data redundancy, or resistance to disk failures. [1]

In general, several different RAID configurations are recognized. These configurations,

called RAID levels, define how data is distributed across multiple drives. Properties of individual levels significantly differ in terms of redundancy and performance. Different schemes are referred to by numbers associated with the given level. We distinguish 7 standard RAID levels: [1]

- **RAID 0** – striping – data are interleaved on all disks, no redundancy – disk failure causes failure of the entire array,
- **RAID 1** – mirroring – identical data are mirrored to all disks, remains operational as long as at least one drive is functioning,
- **RAID 2** – bit-level striping with dedicated Hamming-code parity,
- **RAID 3** – byte-level striping with dedicated parity, each sequential byte is written on a different drive, parity is stored on a dedicated disk,
- **RAID 4** – block-level striping with dedicated parity, each file system block is written on a different drive, parity is stored on a dedicated disk,
- **RAID 5** – block-level striping with distributed parity, same as RAID 4, but parity is split among drives,
- **RAID 6** – block-level striping with double distributed parity, parity information is distributed and stored twice (allowing double disk failure).

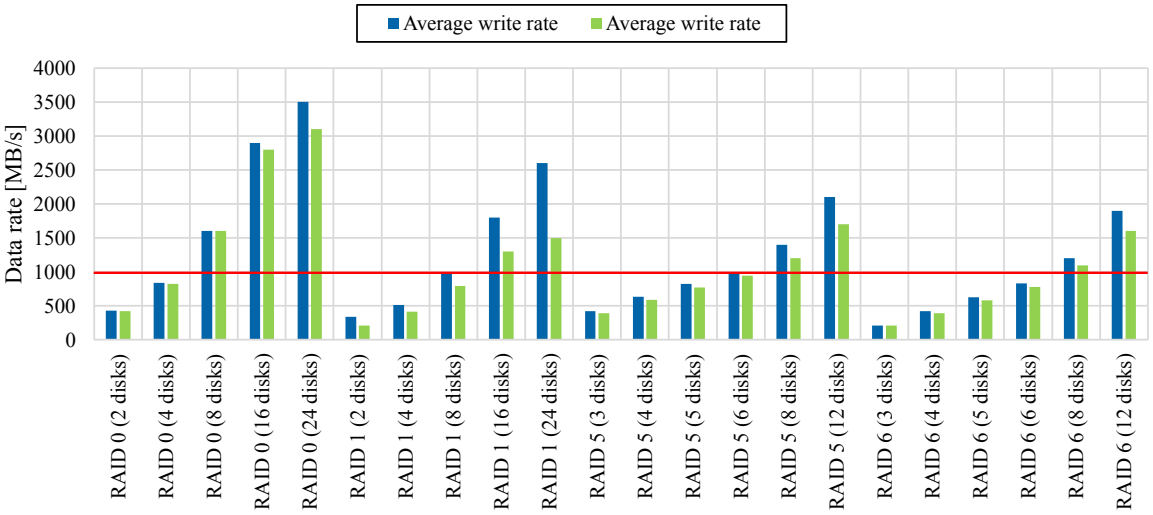


Figure 9: Average read and write data rates of standard RAID configurations

Apart from the standard levels, so-called nested (hybrid) arrays exist. They usually combine two or more RAID levels into a single volume. The most commonly used nested arrays are the following: [3]

- **RAID 00** – striping in two directions,
- **RAID 10** – striping followed by mirroring,
- **RAID 50** – block-level striping with distributed parity followed by simple striping,
- **RAID 60** – block-level striping with double distributed parity followed by simple striping, etc.

The performances of selected nested RAID setups are illustrated in Figure 10. Here, we can notice that these configurations rely on more disks providing superior features, which puts them in a better position than standard configurations.

Choosing the optimal configuration for any given task depends strongly on the desired requirements, e.g., minimal performance, redundancy, and capacity. Plus, individual optimizations are required for every given use case, usually resulting in a trade-off between multiple criteria. In our scenario of storing data generated by the high-energy physics experiment, the final storage system should meet the following requirements:

- sustainable incoming data rate of 1 GB/s,
- sustainable outgoing data rate of 1 GB/s,
- resistance to a single disk failure,
- probability of successful rebuild at least 95 %,
- maximization of the total capacity,
- 24 disk drives in total.

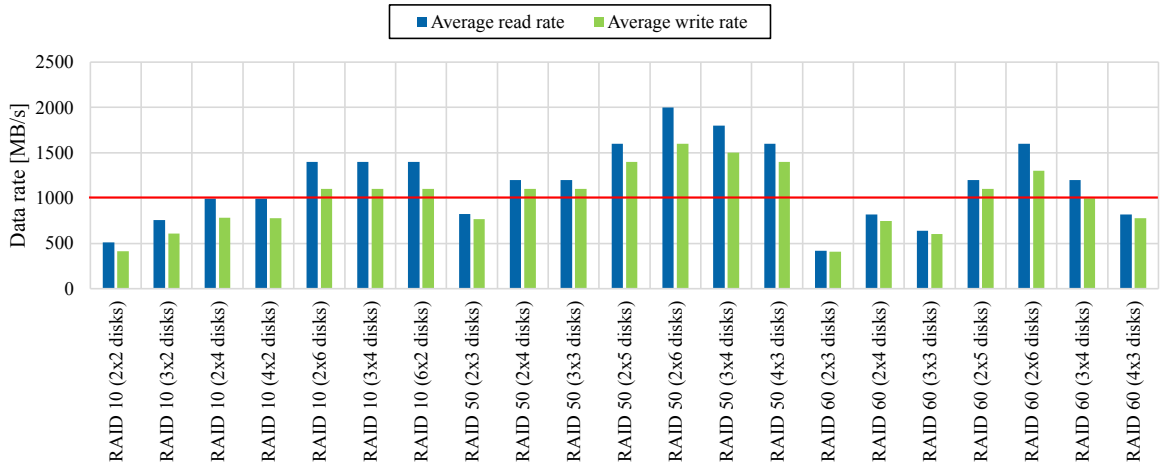


Figure 10: Average read and write data rates of nested RAID configurations

Avoiding parallel access requires at least two volumes. So, it is clear that a single volume can contain only 12 disks at maximum. Together with the requirement for the input and output data rates, we obtain a lower cut for acceptable configurations. Basically, the majority of simple (non-nested) RAID arrays are excluded due to insufficient throughput. In addition, the need for a failing disk resistance effectively eliminates all RAID 0 and RAID 00 configurations that do not provide such a warranty. Eventually, we obtain a limited set of allowed RAID configurations.

3.1 Probability of successful rebuilding

When considering the RAID storage reliability, one of the main concerns is the importance of successful recovery in case of a failing disk. Each disk has a certain amount of broken or corrupted memory cells that are usually caused by an imperfect manufacturing process or wear down of the magnetic material. This metric is called a non-recoverable error rate. Under nominal conditions, this number is negligible in comparison with the total disk capacity and does not have any impact on the disk operation. Plus, the operating system regularly scans drives for broken blocks and excludes them at the file system level.

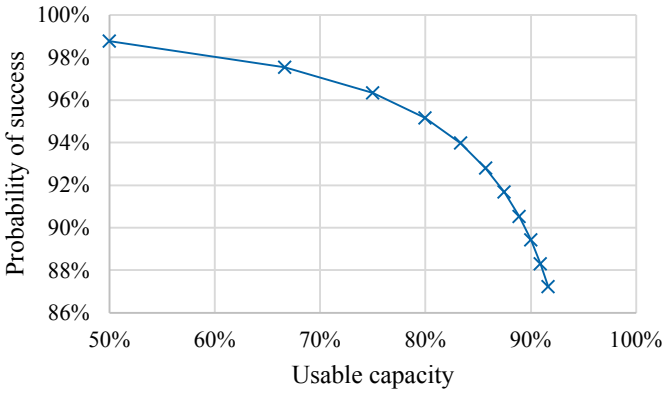


Figure 11: Probability of successful rebuilding of RAID array

However, the non-recoverable error rate can play a significant role when a RAID array is being rebuilt after a broken disk has been replaced. For RAID configurations with parity checks, all information on the remaining disks must be intact to be able to recalculate the original data from parities and to fully reconstruct the broken disk. Since the rebuilding procedure is a resource-intensive process, another disk is likely to fail during this period, or the number of non-recoverable blocks can reach a critical limit.

Consequences strongly depend on the location of broken blocks. If such a block is found in a critical section of the disk (partitioning table, master boot record, operating system files, etc.), it can lead to a corrupted operating system or unreadable data files. Therefore, we must carefully take this aspect into account. The derived formula for the calculation of the probability of successful rebuild (assuming no read error happened) is the following:

$$P_{succ} = \left(1 - \frac{E}{S}\right)^{C \cdot (N-1)}$$

where E is the non-recoverable error rate, S is the data size read for the given error rate, C is the capacity of a single disk, N is the number of disks in RAID span.

Figure 11 shows how the probability of successful rebuilding decreases with bigger RAID spans providing higher capacities. Obviously, there is an inflection point where a trade-off between the probability of successful rebuilding and volume capacity is achieved. This point is unique for different disk models, capacities, RAID configurations, etc. In our requirements, we demand a success probability above 95 % that efficiently eliminates all remaining standard arrays and some nested ones that have a probability of success below this threshold.

3.2 Alternating disk access

As described in the previous sections, when dealing with large data files, the best results can be achieved with a single thread accessing the data. Because our use case requires simultaneous reading and writing, a naive approach is to split the storage into two independent volumes. The first volume would be used only for reading, and the second one would be used only for writing. At some point, the volumes can be swapped, utilizing all

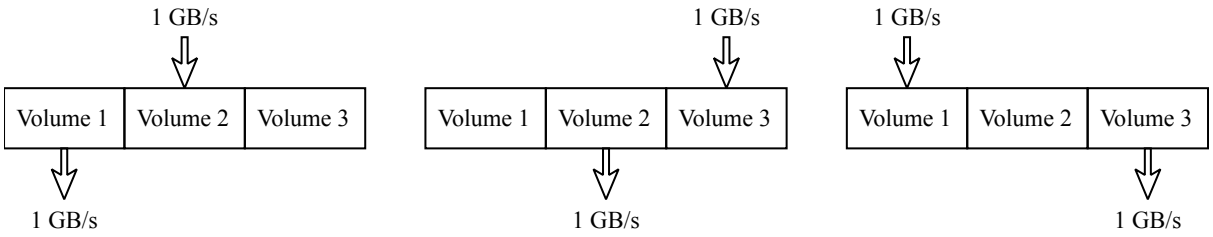


Figure 12: Scheme of alternating access to 3 RAID volumes

disks equally.

However, in case of a single disk failure, the affected volume would be either degraded or inaccessible at all (due to RAID rebuilding). This would result in simultaneous access to the remaining volume, and the data rates would drop significantly. Such an incident is unacceptable and can have a massive impact on the performance of the physics experiment, possibly causing some downtime.

Therefore, we propose to use 3 independent volumes instead of 2. Under the nominal conditions, one volume is read, the second one is written, and the third volume is idle. The writer thread writes to individual volumes in a round-robin manner and iterates through them with half an hour cycle time, i.e., the writer moves to the next volume after 30 minutes. In addition, the writer locks the assigned volume, so there is no simultaneous access from the reading thread. Whereas the reader cannot read from the locked space, it can access the remaining two volumes. Preferably, it chooses the volume that has been unlocked the most recently. In addition, we assume that the readout data rate is slightly higher than the writing data rate on average; otherwise, the storage would overflow. Considering all these aspects, the writer can safely move to the next volume that is not being read by the reader.

Using the described approach, the transitioning between volumes is smooth for the writer as well as for the reader, and parallel disk access does not occur, maximizing the overall performance. Also, in case of a disk failure, the problem is reduced to the usage of two volumes that can still perform at their maximum performance, not causing any downtime to the experiment. As a consequence of using 3 volumes, all nested arrays with more than 8 disks are not acceptable anymore. Therefore, the only remaining and the most suitable RAID configuration is **RAID 50** using 3 volumes per 2 spans per 4 disks ($3 \times 2 \times 4$ configuration).

4 Conclusion

In this paper, we have identified the most optimal data storage solution for a high-energy physics experiment. First, we described the characteristics of a single disk and measured its performance under different types of load using various numbers of threads. We learned that tasks using random read patterns could benefit from parallel access to the disk, mainly due to the optimized head trajectory. When reading large data blocks sequentially, we observed no benefits from accessing the data with multiple threads. On the contrary, the use of many readers has a huge negative impact on the overall disk

performance. The comparison of HDD and SSD disks pointed out that SSDs behave differently. The use of several parallel readers increases the SSD performance regardless of the reading mode (random or sequential).

In our next analysis, we assessed the disk throughput with respect to the disk geometry. Our results show that the performance of the disk decreases with the increasing occupancy; eventually, the throughput dropped down to 48 % of its initial value in our test scenario. Both operations, reading and writing, were affected in the same way roughly by the same factor. We also confirmed our hypothesis by showing that partitions located on the outer disk edge do not suffer from this deficiency.

This was followed by a detailed analysis of various RAID configurations. These tests consisted of read and write measurements using different RAID levels. The first set included only standardized configurations; later, the performance of nested RAID levels was evaluated. We also described the probability of a successful rebuild of a RAID array when a single disk fails. Surprisingly, this probability can drop below 95 % when dealing with spans containing five disks or more. Finally, we proposed an optimized strategy based on alternate disk access, which optimizes the overall performance and increases the reliability of the storage system.

After these thorough measurements, we defined a list of requirements that emerged from our use case in the high-energy physics field. Because the data from these experiments are mostly produced in large data chunks (usually 1 GB or larger), our approach relied on minimizing the number of parallel accesses. In combination with the minimal requested throughputs and other requirements, we gradually eliminated all non-conforming options until there was only one solution left – RAID 50 – 3 volumes per 2 spans (RAID 0) per 4 disks (RAID 5).

References

[1] D. A. Patterson, et al. *A case for Redundant Arrays of Inexpensive Disks (RAID)*. In Proceedings of the 1988 ACM SIGMOD international conference on Management of data. DOI: 10.1145/50202.50214.

[2] S. Huber et al. *Data Acquisition System for the COMPASS++/ AMBER Experiment*. IEEE Transactions on Nuclear Science, vol. 68, no. 8, pp. 1891–1898, 2021. DOI: 10.1109/TNS.2021.3093701.

[3] J. B. Layton. *Intro to Nested-RAID: RAID-01 and RAID-10*. Linux-Mag.com. Linux Magazine. Retrieved 2015-02-01.

[4] H. Grabowski. *Large Spinning Hard Disk Performance Study*. Published 2022-10-19. Retrieved 2023-09-26. Available online: `https://www.nequalsonelifestyle.com/ /2022/10/19/large-spinning-hd-performance-study`

[5] P. Kołaczkowski. *Performance Impact of Parallel Disk Access*. Published 2020-08-24. Retrieved 2023-09-26. Available online: `https://pkolaczk.github.io/ /disk-parallelism`

[6] W. Nicholls. *Toshiba Unveils the World's First 14TB Conventional Hard Drive*. PetaPixel.com Magazine. Published 2017-12-13. Retrieved 2023-09-26.

[7] M. Zemko et al. *Triggerless data acquisition system for the AMBER experiment*. In Proceedings of 41st International Conference on High Energy Physics — PoS(ICHEP2022), Bologna, Italy: Sissa Medialab, 2022-11, p. 248. DOI: 10.22323/1.414.0248.