Mathematisches Forschungsinstitut Oberwolfach

Report No. 55/2021

DOI: 10.4171/OWR/2021/55

Applied Harmonic Analysis and Data Science (hybrid meeting)

Organized by
Ingrid Daubechies, Durham
Gitta Kutyniok, München
Holger Rauhut, Aachen
Thomas Strohmer, Davis

28 November – 4 December 2021

Abstract. Data science has become a field of major importance for science and technology nowadays and poses a large variety of challenging mathematical questions. The area of applied harmonic analysis has a significant impact on such problems by providing methodologies both for theoretical questions and for a wide range of applications in signal and image processing and machine learning. Building on the success of three previous workshops on applied harmonic analysis in 2012, 2015 and 2018, this workshop focused on several exciting novel directions such as mathematical theory of deep learning, but also reported progress on long-standing open problems in the field.

Mathematics Subject Classification (2010): 65Txx, 94Axx, 65K05, 15A52.

Introduction by the Organizers

The workshop Applied Harmonic Analysis and Data Processing was organized by Ingrid Daubechies, Gitta Kutyniok, Holger Rauhut and Thomas Strohmer. This meeting was attended by 57 participants from four continents; 20 of them participated in person and 37 participated virtually.

Data Science encompasses signal and image processing, data processing and machine learning. On the one hand it is a quickly growing field of major importance for science, technology and society and on the other hand it is a very rich source of a large variety of mathematical problems. A major challenge is the ever increasing size and complexity of data and the demand for efficient computational methods for processing such data. Mathematical understanding of the underlying structures and algorithms is highly desired. One of the key drivers for a large

number of big data applications is deep learning. Despite its huge success, however, the mathematical theory of deep learning is still in its infancy. Although a few highly exciting mathematical results could be shown very recently, many open problems remain. This means that the amount of new mathematical challenges arising from the need of data analysis and information processing is enormous, with their solution requiring fundamentally new ideas and approaches, with significant consequences in the practical applications.

Applied Harmonic Analysis provides one key approach towards the problem of efficiently representing, decomposing, processing, and analyzing univariate and multivariate functions and data. Its applications range from theoretical ones such as the decomposition of specific operators to more practical ones such as imaging, machine learning, and inverse problems. Research is typically driven by real-world applications leading to mathematically highly challenging questions, thereby also significantly advancing the mathematical understanding of harmonic analysis itself and in turn impacting the respective application. Significant success has been achieved in the last years, and we exemplarily mention the area of compressive sensing, which has revolutionized the way we approach the collection and analysis of large-scale sparse data such as in high-resolution imaging. This field, which also has roots in other areas such as statistics, optimization and random matrices, has reached a mature state and is nowadays considered a mathematical discipline of its own and has triggered exciting new research directions such as provable non-convex optimization in signal processing and data analysis, emerging mathematical foundations for deep learning, structured dictionary learning, and high-dimensional function reconstruction.

This workshop was a concerted effort to bring together researchers with various backgrounds, including harmonic analysis, optimization, probability theory, approximation theory, machine learning, computer science and electrical engineering. The workshop featured 26 talks, thereof several longer overview talks. Moreover, a session of short presentations of 3 minutes took place on Monday, which we call the 3 Minutes of Fame (following Andy Warhols concept of 15 minutes of fame). This session has meanwhile become a tradition and has proven to be an efficient vehicle to ensure that every participant had the possibility to advertise her/his research. At the same time it is very entertaining for the audience. A large part of the attendees participated, ranging from PhD students to renowned professors, contributing to the success of this session.

Let us mention a few highlights from the program:

- **Mathematical Theory of Deep Learning.** A number of talks reported on progress – but also on intriguing open questions – on the theory of deep learning. Rémi Gribonval talked about the role of sparsity in deep learning and how normalization of weights in the learning process leads to improvements. Helmut Boelcskei provided an information theoretic approach to understanding the limits of deep generative neural networks. For the model problem of low-rank matrix reconstruction Dominik Stöger showed that overparameterization together with small initialization leads

to an implicit bias of gradient descent towards low rank matrices, providing some insights on the somewhat mysterious implicit bias phenomenon in deep learning. In a similar direction, Noam Razin reported on the implicit bias phenomenon in the context of tensor-structured neural networks. Nadav Cohen provided convergence guarantees to global minima for learning deep linear networks via gradient descent, rigorously connecting it with the corresponding gradient flow. Martin Genzel reported that solving inverse problems with deep neural networks may be robust with respect to adversarial noise in contrast to classification with deep neural networks. Felix Krahmer reported on neural network approaches for sparse signal recovery. Youness Boutaib reported on classification with reservoir computing, a particular stochastic recurrent neural networks.

- **Solution of the Strohmer conjecture.** A Gabor system is a discrete set of functions arising as modulations and translations of single function. It is of interest whether such a system with modulation and translation parameters taken from a lattice form a so-called frame with good frame constants. Thomas Strohmer conjectured in 2003 that the hexagonal lattice provides a minimal ratio of the frame constants for the corresponding Gabor system generated by the Gaussian function among all lattices. Markus Faulhuber reported on his solution of this conjecture together with Stefan Steinerberger.
- **Ethical Aspects of Deep Learning.** Rachel Ward started her talk by initiating a discussion on ethical aspects of research on deep learning and data science in general. While the mathematical research is clearly fascinating, deep learning methodology can certainly be abused for tasks that may be questionable, to say the least. In particular, issues of data privacy, automated surveillance and military applications come to mind. One may wonder whether one should do research in this direction at all. Arguments that were discussed include:
  - If researchers in academia stop doing research on deep learning and related topics, only researchers in industry would continue which sets these topics "out of control of academia".
  - Academic education needs to also provide PhD/Master/Bachelor students with fundamentals on ethical aspects (via courses, seminars etc.)
  - We as a research community need to provide tools that can circumvent certain issues. For instance, develop methods that transform data sets in ways that remove privacy information but still allow for efficient training of machine learning models.

  This discussion was certainly open-ended. But the participants felt that it was very important to think about ethical implications of our research.

The organizers would like to take the opportunity to thank MFO for providing support and a very inspiring environment for the workshop. The hybrid setup of the workshop worked very well. For many of the participants who could come

in person to Oberwolfach, it was the first workshop since the beginning of the pandemic where they could actually physically meet and discuss with colleagues. They realized only then how much they had missed this part of science. But also the people who could only participate virtually reported that they very much enjoyed and profited from the workshop.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1641185, "US Junior Oberwolfach Fellows".

Workshop (hybrid meeting): Applied Harmonic Analysis and Data Science

Table of Contents

Abstracts

Rapture of the deep: highs and lows of sparsity in a world of depths

Rémi Gribonval

(joint work with Quoc-Tung Le, Elisa Riccietti, Pierre Stock, Léon Zheng)

Attempting to promote sparsity in deep networks is natural to control their complexity, and can be expected to bring other benefits in terms of statistical significance or explainability. Yet, while sparsity-promoting regularizers are well understood in linear inverse problems, much less is known in deeper contexts, linear or not. We show that, in contrast to the linear case, even the simple bilinear setting leads to surprises: ℓ^1 regularization does not always lead to sparsity [1], and optimization with a fixed support can be NP-hard [2]. We nevertheless identify families of supports for which this optimization becomes easy [2] and well-posed [3], and exploit this to derive an algorithm able to recover multilayer sparse factors with certain prescribed supports [4, 5]. Behind much of the observed phenomena are intrinsic scaling ambiguities in the parameterization of deep linear networks, which are also present in ReLU networks. We conclude with a scaling invariant embedding of such networks [6], which can be used to analyze the identifiability of (the equivalence class of) parameters of ReLU networks from their realization.

References

[1] A. Benichoux, E. Vincent, R. Gribonval, *A fundamental pitfall in blind deconvolution with sparse and shift-invariant priors*, Proc. ICASSP 2013.

[2] Q.T. Le, E. Riccietti, R. Gribonval, *Spurious Valleys, Spurious Minima and NP-hardness of Sparse Matrix Factorization With Fixed Support*, 2021, arXiv:2112.00386.

[3] L. Zheng, E. Riccietti, R. Gribonval, *Identifiability in Two-Layer Sparse Matrix Factorization*, 2021, arXiv:2110.01235.

[4] Q.T. Le, L. Zheng, E. Riccietti, R. Gribonval, *Fast learning of fast transforms, with guarantees*, 2021, https://hal.inria.fr/hal-03438881

[5] L. Zheng, E. Riccietti, R. Gribonval, *Hierarchical Identifiability in Multi-layer Sparse Matrix Factorization*, 2021, arXiv:2110.01230

[6] P. Stock, R. Gribonval, *An Embedding of ReLU Networks and an Analysis of their Identifiability*, to appear in Constructive Approximation.

A convex hierarchy for causal optimization

David Gross

(joint work with Laurens Ligthart, Mariami Gachechiladze)

A *causal structure* is a description of the functional dependencies between random variables. Deciding whether a distribution is compatible with a structure is a practically and fundamentally relevant, yet very difficult problem. Only recently has a general class of algorithms been proposed: These *inflation techniques* associate to any causal structure a hierarchy of increasingly strict compatibility tests, where each test can be formulated as a computationally efficient convex optimization

problem. Remarkably, it has been shown that in the classical case, this hierarchy is *complete* in the sense that each non-compatible distribution will be detected at some level of the hierarchy. An inflation hierarchy has also been formulated for causal structures that allow for the observed classical random variables to arise from measurements on quantum states – however, no proof of completeness of this *quantum inflation hierarchy* has been supplied. We construct a first version of the quantum inflation hierarchy that is provably convergent.

Classical causal models: In the formalization of Ref. [2], causal relationships between classical random variables are modeled using *directed acyclic graphs* (DAGs). Each vertex corresponds to a random variable. Arrows denote causal relationships, in the sense that each variable is taken to be a function of its parents in the graph and independent randomness.

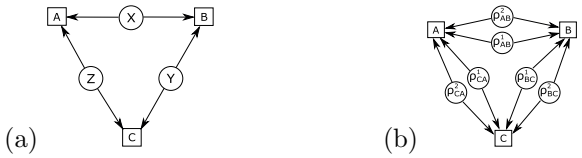


FIGURE 1. Causal structure of the *triangle scenario* (a). Round vertices denote *latent variables* that are not directly accessible, while observed variables are written in squares. Arrows represent causal relations. Panel (b) shows the second level *inflation*.

We are interested in the following causal hypothesis testing problem: Given a joint distribution over the observed random variables and a candidate causal structure, can the distribution be realized in a model that is compatible with the structure? Because there is an infinite set of possible functional relationships, it is a priori not obvious that the causal hypothesis testing problem is even algorithmically decidable.

Recently, hierarchies of convex relaxations for this problem have been developed under the name of *inflation techniques* [3,4]. The high-level idea is to check for the existence of certain *symmetric extensions*. Indeed, assume that a distribution is compatible with a candidate causal structure. One can then define an "inflated" model that involves n independent copies of the hidden variables. This larger model has a number of symmetries: One can exchange a hidden variable from one of the copies with the same hidden variable from another copy, without affecting the distribution (Fig. 1). The n-th level of the hierarchy tests whether such an n-fold inflated model exhibiting all these symmetries exists. In a break-through development, it has been shown that the inflation method is *complete* for the classical causal compatibility problem, in the sense that any incompatible distribution will be detected at some finite level [4].

Quantum causal structures: It is natural to generalize the causal hypothesis testing problem to *quantum causal structures* [5, 8]. The input to the causal hypothesis test is again a directed graph and a joint probability distribution with

one classical variable corresponding to every observed node (Fig. 1). The problem is to decide whether the classical distribution could have arisen from the following process: (i) For each hidden node, prepare a quantum state on as many systems as there are outgoing arrows from that node. Distribute the subsystems along the arrows to the observed nodes. (ii) At each observed node, perform a global measurement on all incoming quantum systems. Assign the result to the observed random variable. Recently, a hierarchy of semidefinite programming (SDP) tests generalizing the inflation technique to the quantum case has been proposed [5].

In Ref. [1], we show completeness of a related SDP hierarchy for the *approximate quantum causal optimization problem*.

Problem 1 (Quantum causal polynomial optimization)**.** *Given a causal structure, a polynomial function f_0 on quantum states, and a countable set $f_1, f_2, \ldots$ of polynomial functions that are non-negative on states compatible with the causal structure. Find*

$$f^\star = \min_\rho f_0(\rho), \quad s.\ t. \quad f_i(\rho) = 0,\ i \geq 1, \quad \rho \text{ compatible with causal structure.}$$

The causal compatibility problem reduces to the optimization one by choosing f_0 to be the 2-norm distance between the observed data P and the one produced by the state. We show:

Theorem 1 (Main Theorem [1])**.** *There is a hierarchy of semidefinite programming relaxations for the quantum causal polynomial optimization problem, which is* complete *in the sense that its optimal values converge to $f^\star$.*

We give a description of such a hierarchy and establish a number of auxiliary results in the course of the work. These include: A proof showing that the arguments in Ref. [6] generalize to give a *quantum de Finetti Theorem* valid for arbitrary C^*-tensor products, not just the minimal tensor product for which it was originally stated. We show that the non-commutative polynomial optimization (NPO) problem treated in Ref. [7] can be interpreted naturally as optimizations over the state space of certain universal C^*-algebras.

References

[1] L.T. Ligthart, M. Gachechiladze, D. Gross. *A convergent inflation hierarchy for quantum causal structures*, arXiv:2110.14659.

[2] J. Pearl, *Causality*, Cambridge university press (2009).

[3] E. Wolfe, R. W. Spekkens and T. Fritz, *The inflation technique for causal inference with latent variables*, Journal of Causal Inference **7**(2) (2019).

[4] M. Navascués and E. Wolfe, *The inflation technique completely solves the causal compatibility problem*, Journal of Causal Inference **8**(1), 70 (2020).

[5] E. Wolfe, A. Pozas-Kerstjens, M. Grinberg, D. Rosset, A. Acín and M. Navascués, *Quantum inflation: A general approach to quantum causal compatibility*, Physical Review X **11**(2), 021043 (2021).

[6] G. Raggio and R. Werner, *Quantum statistical mechanics of general mean field systems*, Helv. Phys. Acta **62**(8) (1989).

[7] S. Pironio, M. Navascués and A. Acin, *Convergent relaxations of polynomial optimization problems with noncommuting variables*, SIAM Journal on Optimization **20**(5), 2157 (2010).
[8] R. Chaves, C. Majenz and D. Gross, *Information–theoretic implications of quantum causal structures*, Nature communications **6**(1), 1 (2015).

Fast binary embeddings using random hyperplane tessellations

Sjoerd Dirksen
(joint work with Shahar Mendelson, Alexander Stollenwerk)

In my talk I will consider the following question. Take independent random hyperplanes with standard Gaussian directions and uniformly distributed shifts. How many hyperplanes are needed to uniformly tessellate a given subset of $\mathbb{R}^n$ with high probability? I will give a generally optimal answer to this question, which surprisingly deviates from the conjectured answer. In the second part of my talk, I will use random hyperplane tessellations to create fast binary embeddings, i.e., fast encodings of a dataset into a minimal number of bits, such that one can quickly query Euclidean distances between the original data points up to a given additive error. For this purpose, I will introduce a computationally efficient structured random matrix, called the double circulant matrix, and will show that it strongly mimics the behavior of a Gaussian matrix. The talk will be based on two forthcoming joint works with Shahar Mendelson and Alexander Stollenwerk.

Fundamental limits of deep generative neural networks

Helmut Bölcskei
(joint work with Dmytro Perekrestenko, Léandre Eberhard)

We show that every d-dimensional probability distribution of bounded support can be generated through deep ReLU networks out of a 1-dimensional uniform input distribution. What is more, this is possible without incurring a cost–in terms of approximation error measured in Wasserstein-distance–relative to generating the d-dimensional target distribution from d independent random variables. This is enabled by a space-filling approach. The construction we propose elicits the importance of network depth in driving the Wasserstein distance between the target distribution and its neural network approximation to zero. Finally, we find that, for histogram target distributions, the number of bits needed to encode the corresponding generative network equals the fundamental limit for encoding probability distributions as dictated by quantization theory.

Optimization and generalization in overparameterized low-rank matrix reconstruction: Small random initialization is akin to spectral learning

Dominik Stöger
(joint work with Mahdi Soltanolkotabi)

Modern machine learning models are typically trained in an overparameterized regime. That is, the number of parameters of the model far exceeds the size of the training data. Due to overparameterization, these models in principle have the capacity to fit any set of labels including pure noise (see, e.g. [2]). Despite this high fitting capacity, these models, which are often trained via first-order methods, generalize well on yet unseen test data.

In this talk, we will focus on overparameterized learning in the context of low-rank reconstruction from a few measurements. More precisely, in our model we assume that we are given observations of the form

$$y_i = \langle A_i, XX^T \rangle = \mathrm{trace}(A_i X X^T) \qquad i = 1, \dots, m, \tag{1}$$

where $\{A_i\}_{i=1}^m \subset \mathbb{R}^{n\times n}$ are known symmetric measurement matrices and $X \in \mathbb{R}^{n\times r_\star}$ is a low-rank matrix to be learned. We consider the non-convex loss function

$$\min_{\bar{U}\in\mathbb{R}^{n\times r}} f(\bar{U}) := \min_{\bar{U}\in\mathbb{R}^{n\times r}} \frac{1}{4m}\sum_{i=1}^m \left(y_i - \langle A_i, \bar{U}\bar{U}^T\rangle\right)^2,$$

where $r \geq r_\star$. In this talk we will be especially interested in the overparameterized regime, i.e. $rn \gg m$ (although our theory will apply for all $r \geq r_\star$.) We can rewrite the loss function into the more compact form

$$\min_{\bar{U}\in\mathbb{R}^{n\times r}} f(\bar{U}) := \min_{\bar{U}\in\mathbb{R}^{n\times r}} \frac{1}{4}\|\mathcal{A}\left(\bar{U}\bar{U}^T - XX^T\right)\|^2, \tag{2}$$

where $\mathcal{A} : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^m$ is the measurement operator defined by $[\mathcal{A}(Z)]_i := \frac{1}{\sqrt{m}}\langle A_i, Z\rangle$. We minimize f via gradient descent, i.e.,

$$\begin{aligned} U_{t+1} &= U_t - \mu\nabla f(U_t) \\ &= U_t + \mu\left[(\mathcal{A}^*\mathcal{A})\left(XX^T - U_tU_t^T\right)\right]U_t. \end{aligned}$$

Here, our initialization U_0 is given via

$$U_0 = \alpha U. \tag{3}$$

The parameter $\alpha > 0$ is referred to as *scale of initialization* and $U \in \mathbb{R}^{n\times r}$ is a random matrix, which is called the shape matrix.

There are two challenges connected with analyzing randomly initialized gradient descent in this setting. Since f is *non-convex* it is a priori not clear whether gradient descent converges to a global optimum or whether it gets stuck in a local minima and/or saddle. This is the first challenge. The second challenge is that of *generalization*. Namely, in the overparameterized scenario the number of parameters is larger than the number of data points i.e. $rn \geq m$. In this case, it can be shown there are infinitely many $\bar{U}$ such that $f(\bar{U}) = 0$, but $\|\bar{U}\bar{U}^T - XX^T\|_F$ is arbitrarily large. That is, even if gradient descent converges to a global optimum,

i.e. $f\left(\bar{U}\right) = 0$, this does not imply that it has found the low-rank solution XX^T or even a point close to it.

In this talk we will address the outlined challenges by discussing the critical role of small random initialization. (To be precise, by small random initialization we mean that the parameter α in (3) is chosen small and that the matrix U in (3) is a Gaussian matrix with i.i.d. entries.)

In this talk, we will discuss the following central insight.

> ***Small random initialization*** *followed by a few iterations of gradient descent behaves akin to spectral initialization.*

Based on this insight, we will derive convergence and generalization guarantees for any $r \geq r_\star$ under the assumption that the scale of initialization $\alpha > 0$ is chosen small enough. We will show that the ground truth signal can be recovered, if the number of samples m is at least at the order of $mr_\star^2$ (under the assumption that the measurement matrices A_i are i.i.d. Gaussian matrices). Moreover, our theory will explain how different choices of r (i.e. varying degrees of overparameterization) impact the trajectory of the gradient descent iterates.

Hence, in this talk, we will give further theoretical support to the empirical observations in [3]. Let us mention that the special case $r = n$ has also been studied in [4]. However, in contrast to our result, the result in [4] requires that the sample size m goes to infinity when the scale of initialization α is decreased to 0.

At the end of the talk, we will discuss a number of open problems. For example, one could aim to extend our results to scenarios where the measurement matrices are more structured such as in matrix completion or blind deconvolution. It is also an interesting open question whether the quadratic dependence of the sample complexity m on $r_\star$ in our results is really necessary or rather an artefact of the proof.

References

[1] D. Stöger, M. Soltanolkotabi *Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction*, accepted in NeurIPS 2021.

[2] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals *Understanding deep learning requires rethinking generalization*, ICLR 2017.

[3] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, N. Srebro *Implicit regularization in matrix factorization*, Advances in Neural Information Processing Systems 2017, 6151-6159.

[4] Y. Li, T. Ma, H. Zhang *Algorithmic Regularization in Over-parameterized Matrix Recovery and Neural Networks with Quadratic Activations*, COLT 2018.

Non-commutative Concentration Inequalities

Afonso Bandeira, March Boedihardjo
(joint work with Ramon van Handel)

Matrix Concentration inequalities such as Matrix Bernstein inequality have played an important role in many areas of pure and applied mathematics [5]. These inequalities are intimately related to the celebrated noncommutative Khintchine inequality of Lust-Piquard and Pisier, which yields a nonasymptotic bound on the spectral norm of general Gaussian random matrices

$$X = \sum_{i=1}^{n} g_i A_i,$$

where g_i are iid standard Gaussian variables and A_i are matrix coefficients.

This bound exhibits a logarithmic dependence on dimension that is sharp when the matrices A_i commute, but often proves to be suboptimal in the presence of noncommutativity. In an Oberwolfach workshop in 2014 [3], Tropp posed the question of whether noncommutativity creates cancellations that can be leveraged to remove this factor (see also [4], [7] and [5] for versions of this question). Subsequently, Bandeira and van Handel [2] obtained sharp bounds for random matrix models with independent entries (with no dimensional dependency in most cases). In a larger range of instances, Tropp [6] improved the dimensional dependence (to a smaller power of the logarithm of the dimension) of noncommutative Khintchine inequality by leveraging non-commutativity of the matrix coefficients.

In this talk we describe our recent work [1], in which we leverage ideas from Free Probability to fully remove this dimensional dependence in a range of instances, yielding optimal bounds in many settings of interest. We make use of the mechanism introduced in [6] to leverage cancellations, together with an interpolation argument that allows us to directly compare the random matrix of interests with its "free"analogue. This argument allows us to also show strong asymptotic freeness (in the sense of Haagerup-Thorbjørnsen) for a remarkably general class of Gaussian random matrices.

As a byproduct we develop matrix concentration inequalities that capture noncommutativity (or, to be more precise, "freeness"), improving over Matrix Bernstein in a range of instances.

References

[1] A. S. Bandeira, M. T. Boedihardjo and R. van Handel, , Matrix Concentration Inequalities and Free Probability, Preprint, arXiv:2108.06312 [math.PR].

[2] A. S. Bandeira and R. van Handel, Sharp nonasymptotic bounds on the norm of random matrices with independent entries, Ann. Prob. Math. Ann. **44** (2016), no. 4, 2478–2506.

[3] Oberwolfach Mini-Workshop: Mathematical Physics meets Sparse Recovery: 2014-04-13 – 2014-04-19, MFO Report 18/2014.

[4] R. I. Oliveira, Sums of random Hermitian matrices and an inequality by Rudelson, Elect. Comm. in Probab. 15 (2010), 203–212.

[5] J. A. Tropp, An introduction to matrix concentration inequalities, Foundations and Trends in Machine Learning, 8:1–230, 2015.

[6] J. A. Tropp, Second-order matrix concentration inequalities, Appl. Comput. Harmon. Anal. **44** (2018), no. 3, 700-736.
[7] R. Vershynin, Spectral norm of products of random and deterministic matrices, Probability Theory and Related Fields, volume 150 (2011),471–509.

Nonlinear Fourier Series

Stefan Steinerberger
(joint work with Raphy Coifman, Hau-tieng Wu)

We discussed a natural way of decomposing a function $F : \mathbb{C} \rightarrow \mathbb{C}$ that is holomorphic in a neighborhood of the unit disk. Our starting point is a fundamental theorem in complex analysis: Blaschke factorization. Any such function F can be decomposed as

$$F = B \cdot G,$$

where B is a Blaschke product, a function of the form

$$B(z) = z^m \prod_{i \in I} \frac{\overline{a_i}}{|a_i|} \frac{z - a_i}{1 - \overline{a_i} z},$$

where $m \in \mathbb{N}_0$ and $a_1, a_2, \cdots \in \mathbb{D}$ are zeroes inside the unit disk $\mathbb{D}$ and G has no roots in $\mathbb{D}$. For $|z| = 1$ we have $|B(z)| = 1$ which motivates the analogy

$$B \sim \text{frequency} \quad \text{and} \quad G \sim \text{amplitude}$$

for the function restricted to the boundary. However, the function G need not be constant: it can be any function that never vanishes inside the unit disk. If F has roots inside the unit disk, then the Blaschke factorization $F = B \cdot G$ is going to be nontrivial (meaning $B \not\equiv 1$ and $G \not\equiv F$). G should be 'simpler' than F because the winding number around the origin decreases.

Unwinding. There is a natural way of iterating Blaschke factorization. G has no zeroes inside the unit disk but the function $G(z) - G(0)$ has at least one root inside the unit disk $\mathbb{D}$ thus has a nontrivial Blaschke factorization $G(z) - G(0) = B_1 G_1$ where B_1 is not constant. Iterating this procedure

$$\begin{aligned} F &= B \cdot G \\ &= B \cdot (G(0) + (G(z) - G(0))) \\ &= B \cdot (G(0) + B_1 G_1) \\ &= G(0) B + B B_1 G_1. \end{aligned}$$

Formally this gives rise to the *unwinding series*

$$F = a_1 B_1 + a_2 B_1 B_2 + a_3 B_1 B_2 B_3 + a_4 B_1 B_2 B_3 B_4 + \dots$$

This unwinding series has a number of desirable properties. One such result, also illustrated in Fig. 1, is monotonicity in the Dirichlet space: if $F \in \mathcal{H}^\infty(\mathbb{D})$

with roots $\{\alpha_j : j \in J\}$ in $\mathbb{D}$ and has the Blaschke factorization $F = B \cdot G$, then

$$\int_{\mathbb{D}} |F'(z)|^2 dz = \int_{\mathbb{D}} |G'(z)|^2 dz + \frac{1}{2} \int_{\partial \mathbb{D}} |G|^2 \sum_{j \in J} \frac{1 - |\alpha_j|^2}{|z - \alpha_j|^2}.$$

The same type of unwinding is possible in other settings. We discuss an analogous factorization on the upper half-space. The suitable replacement of Blaschke products is now given by functions indexed by $\lambda_1, \ldots, \lambda_n \in \mathbb{C}_+$ of the form

$$B(z) = \prod_{k=1}^{n} \frac{z - \lambda_k}{z - \overline{\lambda_k}}, \qquad \text{which satisfy } |B(z)| = 1 \text{ on } \mathbb{R}.$$

We will consider function space $\|\cdot\|_X, \|\cdot\|_Y$ on the set

$$L^2_+(\mathbb{R}) = \left\{ f \in L^2(\mathbb{R}) : \operatorname{supp}(\widehat{f}) \subseteq [0, \infty) \right\}$$

defined by $\psi : [0, \infty] \to [0, \infty]$ which we assume to be a monotonically increasing, differentiable function with $\psi(0) = 0$ and

$$\|f\|_X^2 := \int_0^\infty |\widehat{F}(\xi)|^2 \psi(\xi) d\xi \qquad \text{as well as} \qquad \|f\|_Y^2 := \int_0^\infty |\widehat{F}(\xi)|^2 \psi'(\xi) d\xi.$$

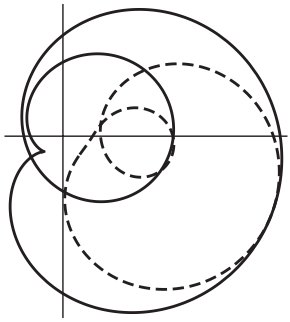


FIGURE 1. $F(e^{it})$ (solid line) given by the cubic polynomial. $G(e^{it})$ (dashed) has the same maximum winding but over a smaller area.

There exists a general convergence result in this setting.

Theorem (see [2]). *If F has roots $\lambda_1, \ldots, \lambda_n \in \mathbb{C}_+$, then we have*

$$\left\| F \prod_{i=1}^{n} \frac{z - \overline{\lambda_k}}{z - \lambda_k} \right\|_X^2 \leq \|F\|_X^2.$$

For the removal of a single root $F(\lambda) = 0$, we have the stronger estimate

$$\left\| F \frac{z - \overline{\lambda}}{z - \lambda} \right\|_X^2 \leq \|F\|_X^2 - (2\Im(\lambda)) \left\| F \frac{z - \overline{\lambda}}{z - \lambda} \right\|_Y^2.$$

Moreover, in the Dirichlet space $\psi(\xi) = \xi$, we even have

$$\left\| F \prod_{i=1}^{n} \frac{z - \overline{\lambda_k}}{z - \lambda_k} \right\|_X^2 \leq \|F\|_X^2 - \int_{\mathbb{R}} |F(x)|^2 \sum_{i=1}^{n} \frac{2\Im(\alpha)}{|x - \lambda_k|^2} dx,$$

where the sum ranges over all roots of F on $\mathbb{C}_+$.

The decomposition has been actively studied in a variety of pure and applied settings. Perhaps the most pressing open question is as follows.

> **Problem.** What can be rigorously established about the convergence properties of the unwinding series? Is it possible to obtain an effective convergence rate for sufficiently nice functions F?

An approach proposed by S. Steinerberger and H. - T. Wu is studied the behavior of one step in the iteration, replacing $G(z)$ by $G(z) - G(0)$ in the special case where G is a random polynomial (see [9]). We conclude with a nice recent observation of Coifman & Peyriere: given

$$B(x) = \prod_{k \geq 1} \frac{x - a_k}{x - \overline{a_k}}$$

we have $B(x) = e^{i\theta(x)}$ where

$$\theta(x) = \sum_{k \geq 0} \sigma \left(\frac{x - \text{Re } a_k}{\text{Im } a_k} \right)$$

and σ is well-known object

$$\sigma(x) = \frac{\pi}{2} + \arctan x.$$

References

[1] R. Coifman and J. Peyriere, Phase unwinding, or invariant subspace decompositions of Hardy spaces, Journal of Fourier Analysis and Applications volume 25, pages 684–695 (2019).

[2] R. Coifman and S. Steinerberger, Nonlinear phase unwinding of functions, Journal of Fourier Analysis and Applications, J. Fourier Anal. Appl. 23 (2017), no. 4, 778–809.

[3] R. Coifman, S. Steinerberger and H.-T. Wu, Carrier frequencies, holomorphy and unwinding, SIAM J. Math. Anal., 49, 4838–4864, (2017).

[4] M. Nahon, Phase Evaluation and Segmentation, Ph.D. Thesis, Yale University, 2000

[5] T. Qian, L.H. Tan, Y.B. Wang, Adaptive Decomposition by Weighted Inner Functions: A Generalization of Fourier Series, Journal of Fourier Analysis and Applications 17, p. 175-190.

[6] T. Qian and L. Zhang, Mathematical theory of signal analysis vs. complex analysis method of harmonic analysis, Appl. Math. J. Chinese Univ, 2013, 28(4): 505-530.

[7] T. Qian, L. Zhang and Z. Li, Algorithm of Adaptive Fourier Decomposition, IEEE Transactions on Signal Processing 59 (2011), p 5899 - 5906.

[8] T. Qian, I. T. Ho, I. T. Leong and Y. B. Wang, Adaptive decomposition of functions into pieces of non-negative instantaneous frequencies, International Journal of Wavelets, Multiresolution and Information Processing, 8 (2010), no. 5, 813-833.

[9] S. Steinerberger and H.-T. Wu, On Zeroes of Random Polynomials and an Application to Unwinding, International Mathematics Research Notices 13, p. 10100–10117 (2021).

[10] G. Weiss and M. Weiss, A derivation of the main results of the theory of Hp-spaces. Rev. Un. Mat. Argentina 20 1962 63-71.

Equivariant machine learning, structured like classical physics

Soledad Villar

(joint work with David W. Hogg, Kate Storey-Fisher, Weichi Yao, Ben Blum-Smith)

There has been enormous progress in the last few years in designing neural networks that respect the fundamental symmetries and coordinate freedoms of physical law. Some of these frameworks make use of irreducible representations, some make use of group convolutions, some use high-order tensor objects or spherical harmonics, and some apply symmetry-enforcing constraints. Different physical laws obey different combinations of fundamental symmetries, but a large fraction (possibly all) of classical physics is equivariant to translation, rotation, reflection (parity), boost (relativity), and permutations. Here we show that it is simple to parameterize universally approximating polynomial functions that are equivariant under these symmetries, or under the Euclidean, Lorentz, and Poincaré groups, at any dimensionality d. The key observation is a simple consequence of classical invariant theory: Nonlinear $\mathrm{O}(d)$-equivariant (and related-group-equivariant) functions can be universally expressed in terms of a lightweight collection of scalars—scalar products and scalar contractions of the scalar, vector, and tensor inputs.

Given a function $f : \mathcal{X} \to \mathcal{Y}$ and a group G acting on $\mathcal{X}$ and in $\mathcal{Y}$ as $\star$ (possibly the action is defined differently in $\mathcal{X}$ and $\mathcal{Y}$). We say that f is:

(1) G-invariant: $f(g \star x) = f(x)$ for all $x \in \mathcal{X},\ g \in G$;

(2) G-equivariant: $f(g \star x) = g \star f(x)$ for all $x \in \mathcal{X},\ g \in G$.

Most physics problems satisfy invariances and equivariances with respect to some group action. Equivariant machine learning aims to learn functions from data, that satisfy those invariances or equivariances by design, with the philosophy that imposing these symmetries provides the right inductive bias for the learning problem. To this end there are many ways to parameterize classes of invariant and equivariant functions. Each of these ways has advantages and disadvantages.

In [1] we provide a complete and computationally tractable characterization of all scalar functions $f : (\mathbb{R}^d)^n \to \mathbb{R}$, and of all vector functions $h : (\mathbb{R}^d)^n \to \mathbb{R}^d$ that satisfy all of the symmetries of classical physics. The groups corresponding to these symmetries are given in Table 1; they act according to the rules in Table 2. The characterization we provide is physically principled: It is based on the fundamental theorem of invariant functions [2] and it is also connected to the symmetries encoded in the Einstein summation rules, a common notation in physics to write expressions compactly but that also allows only equivariant objects to be produced.

Our characterization is based on simple mathematical observations. The first is the First Fundamental Theorem of Invariant Theory for $\mathrm{O}(d)$: *a function of vector inputs returns an invariant scalar if and only if it can be written as a function only of the invariant scalar products of the input vectors* [2, Section II.A.9]. There are similar statements for the Lorentz group $\mathrm{O}(1, d)$ and the rotation group $\mathrm{SO}(d)$. The second observation is that *a function of vector inputs returns an equivariant*

Orthogonal	$\mathrm{O}(d) = \{Q \in \mathbb{R}^{d\times d} : Q^\top Q = Q\,Q^\top = I_d\}$,
Rotation	$\mathrm{SO}(d) = \{Q \in \mathbb{R}^{d\times d} : Q^\top Q = Q\,Q^\top = I_d,\ \det(Q) = 1\}$
Translation	$\mathrm{T}(d) = \{w \in \mathbb{R}^d\}$
Euclidean	$\mathrm{E}(d) = \mathrm{T}(d) \rtimes \mathrm{O}(d)$
Lorentz	$\mathrm{O}(1,d) = \{Q \in \mathbb{R}^{(d+1)\times(d+1)} : Q^\top \Lambda\, Q = \Lambda,\ \Lambda = \mathrm{diag}([1,-1,\ldots,-1])\}$
Poincaré	$\mathrm{IO}(1,d) = \mathrm{T}(d+1) \rtimes \mathrm{O}(1,d)$
Permutation	$\mathrm{S}_n = \{\sigma : [n] \to [n] \text{ bijective function}\}$

TABLE 1. **The groups considered in [1].**

Orthogonal; Lorentz	$Q \star (v_1, \cdots, v_n) = (Q\,v_1, \cdots, Q\,v_n)$
Translation	$w \star (v_1, \cdots, v_n) = (v_1 + w, \cdots, v_k + w, v_{k+1}, \ldots, v_n)$ (where the first k vectors are position vectors)
Euclidean; Poincaré	$(w, Q) \star (v_1, \cdots, v_n) = (Q\,v_1 + w, \cdots, Q\,v_k + w, Q\,v_{k+1}, \cdots, Q\,v_n)$
Permutation	$\sigma \star (v_1, \ldots, v_n) = (v_{\sigma(1)}, \ldots, v_{\sigma(n)})$

TABLE 2. **The actions of the groups on vectors.** For the Euclidean group, the position vectors are positions of points; for the Poincaré group, the position vectors are positions of *events*.

vector if and only if it can be written as a linear combination of invariant scalar functions times the input vectors. In particular, if $h : (\mathbb{R}^d)^n \to \mathbb{R}^d$ of inputs $v_1, \ldots, v_n$ is $\mathrm{O}(d)$ or $\mathrm{O}(1,d)$-equivariant, then it can be expressed as:

$$h(v_1, v_2, \cdots, v_n) = \sum_{t=1}^{n} f_t\Big(\langle v_i, v_j\rangle_{i,j=1}^{n}\Big)\, v_t\ , \tag{3}$$

where f_t can be arbitrary functions, but if h is a polynomial function the f_t can be chosen to be polynomials. In other words, the $\mathrm{O}(d)$ and $\mathrm{O}(1,d)$-equivariant vector functions are generated as a module over the ring of invariant scalar functions by the projections to each input vector. In this expression, $\langle\cdot,\cdot\rangle$ denotes the invariant scalar product, which can be the usual Euclidean inner product, or the Minkowski inner product defined in terms of a metric Λ (see Table 1):

$$\text{Euclidean: } \langle v_i, v_j\rangle = v_i^\top v_j\ , \qquad \text{Minkowski: } \langle v_i, v_j\rangle = v_i^\top \Lambda\, v_j\ . \tag{4}$$

Our work [1] uses the ideas described above, and extension to translations and permutations, to provide simple and universal parameterizations of invariant and equivariant functions. In [3] we apply this scalar-based model to a simple dynamical system, were we show implementations based on this formulation can obtain state-of-the-art numerical performance.

REFERENCES

[1] S. Villar, D. W. Hogg, K. Storey-Fisher, W. Yao, B. Blum-Smith *Scalars are universal: Equivariant machine learning, structured like classical physics*, NeurIPS 2021.
[2] H. Weyl. *The Classical Groups*. Princeton University Press, 1946.

[3] W. Yao, D. W. Hogg, K. Storey-Fisher, S. Villar *A simple equivariant machine learning method for dynamics based on scalars*, NeurIPS Workshop Machine Learning for Physics 2021.

(Non-)uniqueness of phase recovery from Gabor and wavelet transform measurements

Rima Alaifari

(joint work with Francesca Bartolucci, Matthias Wellershoff)

Originating from imaging and audio processing applications, the problem of *phase retrieval* has gained significant attention by the applied harmonic analysis community over the last decades. A common formulation of the problem is the following:

Suppose we consider a Hilbert space $\mathcal{H}$ and a measurement system $(\varphi_\lambda)_{\lambda\in\Lambda} \subset \mathcal{H}$ for an index set $\Lambda \subseteq \mathbb{C}$. What can be said about uniqueness and stability of recovering functions $f \in \mathcal{H}$ from magnitude measurements

$$(|\langle f, \varphi_\lambda\rangle|)_{\lambda\in\Lambda}\,?$$

In its most general formulation, one can study this problem under the assumption that $(\varphi_\lambda)_{\lambda\in\Lambda}$ is a (possibly non-discrete) frame. Dropping the frame condition is not very meaningful, as one would want to consider a measurement system for which any $f \in \mathcal{H}$ can be stably and uniquely determined from $(\langle f, \varphi_\lambda\rangle)_{\lambda\in\Lambda}$ to begin with.

Phase retrieval is a non-linear inverse problem that can be formulated more precisely as follows: Let $\Phi := (\varphi_\lambda)_{\lambda\in\Lambda}$ be a frame for $\mathcal{H}$ and define a metric that accounts for inevitable trivial ambiguities, i.e.

$$d_{\mathcal{H}}(f, g) := \inf_{\tau\in\mathcal{S}^1} \|f - \tau g\|,$$

where $\mathcal{S}^1$ is the unit circle in $\mathbb{C}$. Furthermore, let

$$\mathbb{P}_{\mathbb{C}}\mathcal{H} := \mathcal{H}/\mathcal{S}^1$$

be the according quotient space. Then, phase retrieval amounts to the inversion of the operator

$$\mathcal{A}_\Phi : \mathbb{P}_{\mathbb{C}}\mathcal{H} \to \mathbb{R}_+^\Lambda, \quad f \mapsto (|\langle f, \varphi_\lambda\rangle|)_{\lambda\in\Lambda}.$$

In other words, the goal is to recover signals up to a global phase factor, since it is not possible to distinguish f from τf, $\tau \in \mathcal{S}^1$, when only phaseless measurements are acquired.

We have recently studied the question of uniqueness for phase retrieval when the measurement system is a Gabor or wavelet frame. This consideration is motivated by applications such as *ptychographic imaging* and the reconstruction problem in the *phase vocoder,* an application in audio processing. First, let us define the continuous transforms. The windowed (or short-time) Fourier transform (STFT) of a function $f \in L^2(\mathbb{R})$ for a suitable window g is given by

$$V_g f(x, y) := \int_{\mathbb{R}} f(t)\overline{g(t - x)}e^{-2\pi ity}dt, \quad x, y \in \mathbb{R}.$$

The continuous wavelet transform of a function $f \in L^2(\mathbb{R})$ with respect to a suitable wavelet ψ is defined as

$$W_\psi f(b,a) := \frac{1}{a}\int_{\mathbb{R}} f(t)\overline{\psi\left(\frac{t-b}{a}\right)}dt, \quad a \in \mathbb{R}_+, b \in \mathbb{R}.$$

We may then consider the following setup for phase retrieval:

Given magnitudes of the STFT or wavelet transform in the form

$$\{|V_g f(x,y)|\}_{(x,y)\in\Lambda}, \quad \Lambda \subseteq \mathbb{R}^2$$

or

$$\{|\mathcal{W}_\psi f(b,a)|\}_{(b,a)\in\Lambda}, \quad \Lambda \subseteq \mathbb{R}\times\mathbb{R}_+,$$

respectively, when is phase retrieval uniquely solvable? In particular, is it possible to identify windows g (or wavelets ψ), and choices of index sets Λ and subspaces $\mathcal{M} \subseteq L^2(\mathbb{R})$ such that the phase retrieval operator $\mathcal{A}_\Phi : \mathbb{P}_\mathbb{C}\mathcal{M} \to \mathbb{R}_+^\Lambda$ is injective? Especially, can we choose Λ to be discrete?

In recent work, we have given a positive answer for wavelets [2], and a positive [1] as well as a negative answer [3] for the STFT. The two positive results are both given for $\mathcal{M}$ the Paley-Wiener space defined as

$$\mathrm{PW}_\Omega = \{f \in L^2(\mathbb{R}) : \ \mathrm{supp}\widehat{f} \subseteq [-\Omega,\Omega]\}.$$

Uniqueness of wavelet phase retrieval in the Paley-Wiener space. It is possible to state a uniqueness result for wavelet phase retrieval when the wavelet has a finite number of *vanishing moments.* If ψ has exactly ℓ vanishing moments, then the function $a^{-\ell}W_\psi f(\cdot,a)$ converges to f in the L^2-norm. This allows to give a uniqueness result for real-valued functions $f \in \mathrm{PW}_\Omega$ from measurements

$$\left|W_\psi f\left(\frac{m}{4\Omega}, a_k\right)\right|, \quad m,k \in \mathbb{N},$$

where $(a_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ is a sequence converging to zero. The main ingredients here being that PW_Ω is a reproducing kernel Hilbert space, the fact that analytic functions real-valued on the real line are uniquely determined (up to sign) by their magnitudes [5] and Shannon's sampling theorem.

Uniqueness of Gabor phase retrieval in the Paley-Wiener space. The so-called *ambiguity function relation* of time-frequency (TF) analysis relates the magnitude of the STFT to the ambiguity function of the signal:

$$\mathcal{F}\left(|V_g f|^2\right)(y,-x) = \mathcal{A}f(x,y)\cdot\overline{\mathcal{A}g(x,y)},$$

where $\mathcal{F}$ denotes the (two-dimensional) Fourier transform and $\mathcal{A}f(x,y) = e^{i\pi xy}V_f f(x,y)$ denotes the ambiguity function of f. The formula states that a sufficient condition for $\mathcal{A}f$ to be uniquely recoverable from the STFT magnitudes is that the ambiguity function of the window g has no zeros. It is easy to show that $\mathcal{A}f$ determines f up to a global phase factor.

The most prominent example of a window with non-vanishing ambiguity function is the Gauss function. A less obvious example is the one-sided exponential [4].

Note that in our context, it is sufficient that the ambiguity function only vanishes on sets of measure zero, so that the Hermite windows can also be included.

We then searched for sufficient conditions on the ambiguity function of the window, when the signal class to be reconstructed is the Paley-Wiener space. We found that for the recovery of real-valued signals, it is sufficient to have that the ambiguity function of the window g is non-zero almost everywhere on a finite line segment of the TF plane, i.e. to have

$$\mathcal{A}g(0,y) \neq 0, \quad \text{for a.a. } y \in (-2\Omega, 2\Omega).$$

When the signals are complex-valued, it suffices that $\mathcal{A}g$ is non-zero almost everywhere on two finite line segments: uniqueness of phase retrieval can be guaranteed if there is a $c \in (0, \frac{1}{2\Omega}]$ such that

$$\mathcal{A}g(0,y) \neq 0, \ \mathcal{A}g(c,y) \neq 0, \quad \text{for a.a. } y \in (-2\Omega, 2\Omega).$$

In the real-valued setting (i.e. recovery of real-valued signals f from $|V_g f|$) one can also state a uniqueness result from samples of $|V_g f|$, if the window g is such that its Fourier transform is non-zero almost everywhere on $(-\Omega, \Omega)$. This includes the Gauss function.

Non-uniqueness of Gabor phase retrieval in $L^2(\mathbb{R})$. For a while, it has been an open question whether for phase retrieval from STFT magnitudes there is a *critical density,* i.e. whether uniqueness of phase retrieval can be guaranteed when the Gabor frame is sufficiently oversampled. In light of the existing uniqueness result from STFT magnitudes *without sampling* when the window is the Gauss function, it is particularly interesting to study this choice of window function. We answer this in the negative: when the signal class to be reconstructed is not further restricted, i.e. when $\mathcal{M} = L^2(\mathbb{R})$, then for any lattice $\Lambda = R_\alpha(a\mathbb{Z} \times b\mathbb{Z}) + \lambda$ there exist functions $f_1, f_2 \in L^2(\mathbb{R})$ such that for $\varphi(t) = e^{-\pi t^2}$,

$$\begin{aligned} |V_\varphi f_1(x,y)| &= |V_\varphi f_2(x,y)|, \quad (x,y) \in \Lambda, \text{ but} \\ d_{L^2(\mathbb{R})}(f_1, f_2) &\neq 0. \end{aligned}$$

Here, R_α denotes rotation by α and $\lambda \in \mathbb{C}$ a shift in the TF plane. We note that these counterexamples have a simple explicit form. In the case that Λ is a rectangular lattice, they can even be chosen to be real-valued.

References

[1] R. Alaifari, M. Wellershoff, *Uniqueness of STFT phase retrieval for bandlimited functions*, Applied and Computational Harmonic Analysis **50** (2021), 34–48.

[2] R. Alaifari, F. Bartolucci, M. Wellershoff, *Phase retrieval of bandlimited functions for the wavelet transform*, arXiv:2009.05029.

[3] R. Alaifari, M. Wellershoff, *Phase retrieval from sampled Gabor transform magnitudes: Counterexamples*, arXiv:2010.01078.

[4] K. Gröchenig, P. Jaming, E. Malinnikova, *Zeros of the Wigner distribution and the short-time Fourier transform*, Revista Matemática Complutense **33** (2020), 723–744.

[5] G. Thakur, *Reconstruction of bandlimited functions from unsigned samples*, Journal of Fourier Analysis and Applications **17.4** (2011), 720–732.

Continuous vs. Discrete Optimization of Deep Neural Networks

Nadav Cohen
(joint work with Omer Elkabetz)

The success of deep neural networks is fueled by the mysterious properties of gradient-based optimization, namely, the ability of (variants of) gradient descent to minimize non-convex training objectives while exhibiting tendency towards solutions that generalize well. Vast efforts are being directed at mathematically analyzing this phenomenon, with existing results typically falling into one of two categories: continuous or discrete. Continuous analyses usually focus on gradient flow (or variants thereof), which corresponds to gradient descent (or variants thereof) with infinitesimally small step size. Compared to their discrete (positive step size) counterparts, continuous settings are oftentimes far more amenable to theoretical analysis (*e.g.* they admit use of the theory of differential equations), but on the other hand are stylized, and disregard the critical aspect of computational efficiency (number of steps required for convergence). Works analyzing gradient flow over deep neural networks either accept the latter shortcomings (see for example [1–3]), or attempt to reproduce part of the results via completely separate analysis of gradient descent (*cf.* [4–6]). The extent to which gradient flow represents gradient descent is an open question in the theory of deep learning.

The presented work formally studies the foregoing question. Viewing gradient descent as a numerical method for approximately solving the initial value problem corresponding to gradient flow, we turn to the literature on numerical analysis, and invoke a fundamental theorem concerning the approximation error. The theorem implies that in general, the match between gradient descent and gradient flow is determined by the curvature around gradient flow's trajectory. In particular, the "more convex" the trajectory, *i.e.* the larger the (possibly negative) minimal eigenvalue of the Hessian is around the trajectory, the better the match is guaranteed to be. We show that when applied to deep neural networks (fully connected as well as convolutional) with homogeneous activations (*e.g.* linear, rectified linear or leaky rectified linear), gradient flow emanating from near-zero initialization (as commonly employed in practice) follows trajectories that are "roughly convex," in the sense that the minimal eigenvalue of the Hessian along them is far greater than in arbitrary points in space, particularly towards convergence. This implies that over deep neural networks, gradient descent with moderately small step size may in fact be close to its continuous limit, *i.e.* to gradient flow. We exemplify an application of this finding by translating an analysis of gradient flow over deep linear neural networks into a convergence guarantee for gradient descent. The guarantee we obtain is, to our knowledge, the first to ensure that a conventional gradient-based algorithm optimizing a deep (three or more layer) neural network of fixed (data-independent) size efficiently converges to global minimum *almost surely* under random (data-independent) near-zero initialization.

We corroborate our theoretical analysis through experiments with basic deep learning settings, which demonstrate that reducing the step size of gradient descent often leads to only slight changes in its trajectory. This confirms that, in basic settings, central aspects of deep neural network optimization may indeed be captured by gradient flow. Recent works (*e.g.* [7–9]) suggest that by appropriately modifying gradient flow it is possible to account for advanced settings as well, including ones with momentum, stochasticity and large step size. Encouraged by these developments, we hypothesize that the vast bodies of knowledge on continuous dynamical systems, and gradient flow in particular (see, *e.g.*, [10, 11]), will pave way to unraveling mysteries behind deep learning.

The main contributions of this work are: *(i)* we conduct the first formal study for the discrepancy between continuous and discrete optimization of deep neural networks; *(ii)* we demonstrate the use of *generic* mathematical machinery for translating a continuous non-convex convergence result into a discrete one; *(iii)* to our knowledge, the discrete result we obtain forms the first guarantee of random (data-independent) near-zero initialization *almost surely* leading a conventional gradient-based algorithm optimizing a deep (three or more layer) neural network of fixed (data-independent) size to efficiently converge to global minimum; *(iv)* the fundamental theorem (from numerical analysis) we employ is seldom used in machine learning contexts and may be of independent interest; and *(v)* we provide empirical evidence suggesting that gradient descent over simple deep neural networks is often close to gradient flow.

References

[1] Saxe, Andrew M and McClelland, James L and Ganguli, Surya *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks* International Conference on Learning Representations 2014

[2] Arora, Sanjeev and Cohen, Nadav and Hazan, Elad, *On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization*, International Conference on Machine Learning 2018

[3] Razin, Noam and Cohen, Nadav, *Implicit Regularization in Deep Learning May Not Be Explainable by Norms*, Advances in Neural Information Processing Systems 2020

[4] Ji, Ziwei and Telgarsky, Matus, *Gradient descent aligns the layers of deep linear networks*, International Conference on Learning Representations 2019

[5] Du, Simon S and Hu, Wei and Lee, Jason D, *Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced*, Advances in Neural Information Processing Systems 2018

[6] Arora, Sanjeev and Cohen, Nadav and Golowich, Noah and Hu, Wei, *A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks*, International Conference on Learning Representations 2019

[7] Barrett, David GT and Dherin, Benoit, *Implicit gradient regularization*, International Conference on Learning Representations 2021

[8] Kunin, Daniel and Sagastuy-Brena, Javier and Ganguli, Surya and Yamins, Daniel LK and Tanaka, Hidenori, *Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics*, International Conference on Learning Representations 2021

[9] Smith, Samuel L and Dherin, Benoit and Barrett, David GT and De, Soham, *On the origin of implicit regularization in stochastic gradient descent*, International Conference on Learning Representations 2021

[10] Glendinning, Paul, *Stability, instability and chaos: an introduction to the theory of nonlinear differential equations*, Cambridge university press 1994
[11] Ambrosio, Luigi and Gigli, Nicola and Savaré, Giuseppe, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media

Which neural networks can be computed? - Expressivity meets Turing in Deep Learning

Laura Thesing
(joint work with Anders C. Hansen)

1. Motivation

Deep learning with neural networks celebrates success on a large variety of tasks. The performance was long believed impossible. Popular tasks are image classification, speech recognition, scene detection, inverse problems and PDEs. However, besides the impressive results, we also learn more about the inherent instabilities to small perturbations of the input. The instabilities are shown with adversarial examples. Especially for image classification, there exists a large variety of different attack types. We aim to shed further light on this issue by combining the findings from expressivity theory of neural networks with their computability properties. In this talk, we present a framework for the extension and a first result. Moreover, we discuss the open questions which answers will allow us to give further insight into how to build stable and accurate networks.

Expressivity theory for neural networks dates back to the *universal approximation theorem* [1] and has then been extended to different architecture types as well as smoothness properties of the functions that are approximated. We have a good understanding of which functions can be approximated with neural networks. Given these insights, we ask three further questions which relate the results to their computability.

(1) Is the ground truth a sensible problem?
(2) Are the neural networks in the approximation results computable?
(3) Does there exist an algorithm that maps from (noisy) samples to the neural network approximation?

Finally, one can also ask how can we ensure that the method we chose acts as expected?

2. Expressivity meets Computability

These questions and their motivation are the main part of the talk. We consider the computability in terms of the Turing model and the accuracy in the worst-case scenario with the $\|\cdot\|_\infty$-norm. This can in the future also be extended to average case analysis with for example the $\|\cdot\|_2$-norm. We also had interesting discussions on different choices of the computability model, which are now part of further considerations.

1. The ground truth: We ask if the task, for example, a classification task, is a sensible problem. Informally speaking, the first thought should always be if it is plausible to assume that there is a meaningful structure in the data. Neural networks have a very large representative power and are therefore also able to fit random labels on image data sets [2]. The fitting does not take longer than with meaningful data. Hence, the pure convergence of a learning algorithm does not tell if the underlying problem is sensible nor if the approximation is correct. From a computer science perspective, we ask if the problem relates to a computable function. If the ground truth is already unstable we cannot expect a good approximation to become stable. Moreover, if the underlying function is already non-computable we will see that an accurate approximation with neural networks is impossible. Hence, the understanding of the ground truth is of very high importance but unfortunately also very hard to analyse.

2. The neural networks: This part is already been taken into account in the expressivity literature [3]. Here, it is being controlled that the weights of the network are rational numbers and therefore computable. Moreover, the most commonly used activation function the $ReLU(x) = \max\{0, x\}$ is also a computable function. And hence the composition of computable functions is again computable.

3. The approximating mapping: The core part of every learning algorithm is the mapping from potentially noisy point samples to the approximating network. Here, often a minimization problem with a loss function is solved. We do not take this restriction into account but ask if there is any algorithm that can compute the approximating network. The existence of such an algorithm is directly related to the computability of the ground truth. To make the statement more precise we introduce the computability from point samples. A family of functions $\mathcal{F}$ with $f : [0, 1] \to \mathbb{R}$ is *n-computable from point samples* for $n \in \mathbb{N}$ if there exists an algorithm Γ, which takes as input point samples of f on the dyadic grid and a computable point $x \in \mathbb{R}$, where the function is evaluated. Then Γ outputs an approximation y to the evaluation of $f(x)$ such that $|y - f(x)| < 2^{-n}$. If this holds for all $n \in \mathbb{N}$ we call $\mathcal{F}$ ∞-computable from point samples. This notion allows having finite accuracy of the approximation which is usually sufficient and also takes into account where the information is taken from. We then get for $n \in \mathbb{N}$, a family of functions $\mathcal{F}$ is n-computable from point samples if and only if there exists an algorithm $\Gamma_{\mathcal{F},n}$ which computes for all $f \in \mathcal{F}$ a neural network $\Psi_{f,n}$ that approximates f up to an error of 2^{-n}, i.e.

$$\|\Psi_{f,n} - f\|_\infty \leq 2^{-n}$$

from information of point samples of f.

3. Conclusion, Outlook and Open Questions

We have seen that we can relate the computability from point samples to the existence of an algorithm that gives an approximation with computable neural networks. Hence, all computable problems are also computable with neural networks. This underlines the immense power of neural networks. One may ask why we then still see the instabilities in practice. Two possible reasons are that the

sampling complexity is essential. Our results are constructive and the sampling complexity is much larger than the number of samples used in practice. The other possibility is that the learning algorithm that is mostly used is not the algorithm that can find a good approximation. To get further insight into this question we continue this work with the analysis of different expressivity results with respect to their computability. We then aim to consider also a more restrictive family of functions to reduce the sampling complexity. Finally, the weights in the construction are very large and we aim to reduce them as well with more sophisticated architecture choices.

In the long run, we want to get a good understanding of the ground truth, especially in image classification to understand which problems are computable and sensible. In this line, a measure that allows us to investigate both the stability and accuracy of the network approximation is of high interest.

References

[1] K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators*, Neural networks, **2** (1989), 359–366.
[2] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning (still) requires rethinking generalization*, Communications of the ACM 64 **3** (2021) 107–115.
[3] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, *Optimal approximation with sparsely connected deep neural networks*, SIAM Journal on Mathematics of Data Science, 1 **1** (2019) 8–45.

Norm bounds for a scattering transform on graphs

Bernhard G. Bodmann
(joint work with Íris Emilsdóttir)

Introduction

In the present work, we examine signals on graphs whose structure is motivated by time series analysis. In typical models for time series data, the values that occur closely together in time show a stronger dependence than observations that are spaced further apart [9]. However, this property may need to be modified to explain recurring patterns, such as daily or weekly periodicities in traffic intensities. We assume that the underlying periodicities are known and encoded in a graph structure, where neighboring vertices are immediate successors in time or related by a shift in time corresponding to a period of the observed process. Based on the graph structure, one may devise a type of scattering transform in the spirit of Mallat's method to generate feature vectors with convolutional networks in a non-adaptive way [1, 7]. This has been done by Zou and Lehrman [11] based on graph wavelets [4], see also [2]. Here, we pursue a parallel strategy that is based on heat kernels.

Preliminaries

An oriented graph (V, E) is described by a vertex set V and an edge set E, for which E contains *ordered* pairs of vertices. When considering a directed graph, we also speak of an edge without orientation when passing from $(i, j) \in E$ to $\{i, j\}$. Two edges are adjacent if they have a vertex in common. A graph is connected if any two vertices in V appear in a sequence of vertices such that each pair of consecutive elements in this sequence forms an edge. A directed graph is weakly connected if any two vertices appear in a sequence of adjacent edges without orientation.

The Hilbert space $\ell^2(V)$ is the space of all real-valued functions $f : V \to \mathbb{R}$, equipped with the canonical inner product that associates $\langle f, g\rangle = \sum_{j\in V} f(j)g(j)$ with $f, g \in \ell^2(V)$. The standard graph Laplacian Δ is the self-adjoint operator corresponding to the quadratic form defined by $Q(f) = \sum_{i,j\in E} |f(i) - f(j)|^2$ for $f \in \ell^2(V)$. Given a directed graph (V, E) and a function $a : E \to \mathbb{R}$, we let Δ_a be the operator on $\ell^2(V)$ corresponding to $Q_a(f) = \sum_{(i,j)\in E} |e^{a(i,j)} f(i) - f(j)|^2$. In this context, we call the function a a connection. Finally, for two functions w, a on the edge set of a directed graph with w assuming only strictly positive values, we let $\Delta_{w,a}$ be defined via $Q_{w,a}(f) = \sum_{(i,j)\in E} w(i, j)|e^{a(i,j)} f(i) - f(j)|^2 = -\langle \Delta_{w,a} f, f\rangle$. We then say that the edges are weighted by w. If there is $\phi : V \to \mathbb{R}$ such that for each $(i, j) \in E$, $a(i, j) = \phi(j) - \phi(i)$, then we say that a is a gradient function.

Main results

With the help of the Laplacian $\Delta_{w,a}$, we define a cascade of transforms.

Definition. Let (V, E) be a directed graph with weights $w : E \to \mathbb{R}^+$, connection $a : E \to \mathbb{R}$ and Laplacian $\Delta_{w,a}$, $\langle \Delta_{w,a} f, f\rangle = -\sum_{\langle i,j\rangle} w_{i,j}|e^{a(i,j)} f(i) - f(j)|^2$. For $\epsilon > 0$, $f : V \to \mathbb{R}$, let $S_0(f) = f$, we inductively set for $m \in \mathbb{N}$

$$S_m(f) = (I - (I - \epsilon\Delta_{w,a})e^{\epsilon\Delta_{w,a}})^{1/2}|S_{m-1}(f)|$$

and

$$T_m(f) = (I - \epsilon\Delta_{w,a})^{1/2} e^{\frac{\epsilon}{2}\Delta_{w,a}} |S_{m-1}(f)|\,.$$

Here, for any function g on V, $|g| = \max\{g, -g\}$.

It is a direct consequence of this definition that the norm of a signal is preserved under this transform.

Proposition. Let (V, E), w, a, and $\Delta_{w,a}$, S_m and T_m be as above, then for $N \in \mathbb{N}$,

$$\|f\|^2 = \|S_N(f)\|^2 + \sum_{m=1}^{N} \|T_m(f)\|^2$$

Next, we observe that if a is a gradient function, then the functions in the kernel of the Laplacian saturate the norm of T_1.

Proposition. Let (V, E) be a weakly connected, directed graph, w, a, and $\Delta_{w,a}$, S_1 and T_1 as above, and there is $\phi : V \to \mathbb{R}$ such that $a(i,j) = \phi(j) - \phi(i)$, then $\|T_1(f)\| = \|f\|$ if and only if there is $c \in \mathbb{R}$ and for each $i \in V$, $f(i) = ce^{\phi(i)}$.

More generally, we study the behavior of the norm when applying S_1 or T_1, with a similar motivation as in [10]. Here, we relate the norm to the Rayleigh quotient of the Laplacian. Because of the Parseval-type identity $\|f\|^2 = \|S_1(f)\|^2 + \|T_1(f)\|^2$, it is enough to investigate T_1.

Theorem. Let (V, E) be a weakly connected, directed graph with weights w, connection a, and Laplacian $\Delta_{w,a}$ such that λ_1 is the first non-zero eigenvalue of $-\Delta_{w,a}$, and $\lambda_{\max}$ the maximal eigenvalue of $-\Delta_{w,a}$. If $\epsilon > 0$, $0 \le \rho \le \lambda_1$, and

$$-\langle \Delta_{w,a} f, f\rangle = \rho \|f\|^2$$

then T_1 as defined above satisfies

$$\|T_1(f)\|^2 \ge \big[(1 - \frac{\rho}{\lambda_1}) + (1 + \epsilon\lambda_{\max})e^{-\epsilon\lambda_{\max}} \frac{\rho}{\lambda_{\max}}\big] \|f\|^2$$

and

$$\|T_1(f)\|^2 \le \big[(1 - \frac{\rho}{\lambda_{\max}}) + (1 + \epsilon\lambda_1)e^{-\epsilon\lambda_1} \frac{\rho}{\lambda_1}\big] \|f\|^2 .$$

Corollary. If the assumptions of the preceding theorem hold, $f \in \ell^2(V) \setminus \{0\}$, and $\epsilon > 0$ is sufficiently small so that $(1 + \epsilon\lambda_1)e^{-\epsilon\lambda_1}\lambda_{\max} > \lambda_1$ then

$$\frac{\|T_1(f)\|^2/\|f\|^2 - 1}{(1 + \epsilon\lambda_1)e^{-\epsilon\lambda_1}/\lambda_1 - 1/\lambda_{\max}} \le \rho .$$

Estimating the Rayleigh quotient is useful when it is used as a statistic to infer whether an observed graph signal is consistent with a stochastic model for it. This will be pursued in an application to traffic counts in forthcoming work.

References

[1] J. Bruna and S. Mallat, *Invariant scattering convolution networks*, IEEE Trans. Pattern Anal. Mach. Intell. **35** (8) (2013), 1872-1886.

[2] R. R. Coifman and M. Maggioni, *Diffusion wavelets*, Applied and Computational Harmonic Analysis **21**(1) (2006), 53–94.

[3] F. Gama, A. Ribeiro, and J. Bruna, *Diffusion Scattering Transforms on Graphs*, International Conference on Learning Representations, 2019, Paper 110.

[4] D. K. Hammond, P. Vandergheynst, and R. Gribonval, *Wavelets on graphs via spectral graph theory*, Applied and Computational Harmonic Analysis **30**(2) (2011), 129–150.

[5] A. Sandryhaila and J. M. F. Moura, *Discrete signal processing on graphs*, IEEE Trans. Signal Proc. **61** (2013), 1644–1656.

[6] A. Sandryhaila and J. M. F. Moura, *Discrete signal processing on graphs: Frequency analysis*, IEEE Trans. Signal Proc. **62** (2014), 3042–3054.

[7] Stéphane Mallat, *Group invariant scattering*, Communications on Pure and Applied Mathematics **65** (10) (2012), 1331–1398.

[8] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, *Graph signal processing: Overview, challenges, and applications*, Proceedings of the IEEE **106** (2018), 808–828.

[9] M. B. Priestley, *Spectral Analysis and Time Series*, vols. 1 & 2, Academic Press, London, 1981.

[10] T. Wiatowski, P. Grohs, H. Bölcskei, *Energy propagation in deep convolutional neural networks*, IEEE Trans. Inform. Theory **64** (7) (2017), 4819–4842.
[11] D. Zou and G. Lerman, *Graph convolutional neural networks via scattering*, Applied and Computational Harmonic Analysis, **49**(3) (2020), 1046-1074.

Gaussian lattice sums

Markus Faulhuber
(joint work with Laurent Bétermin, Stefan Steinerberger)

1. Definition and Main Result

In [2], we characterize optimizers for a variational problem with applications in various fields. Let Λ be a lattice in $\mathbb{R}^2$ and consider the function

$$E_\Lambda(z;\alpha) = \sum_{\lambda\in\Lambda} e^{-\pi\alpha|\lambda+z|^2} \qquad z\in\mathbb{R}^2,\ \alpha>0. \tag{1}$$

The function $E_\Lambda(z;\alpha)$ is simply the sum of (scaled) Gaussians centered at points given by a (shifted) lattice: it may thus be understood as the two-dimensional analogue of a Jacobi theta function. Given the fundamental nature of this object, the function $E_\Lambda(z;\alpha)$ naturally arises in many different areas of mathematics. In [2], we are concerned with minimizing and maximizing the function $E_\Lambda(z;\alpha)$. The canonical candidate for solving the variational problem is the hexagonal lattice;

$$\Lambda_2 = \sqrt{\tfrac{2}{\sqrt{3}}}\begin{pmatrix} 1 & \frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2}\end{pmatrix}\mathbb{Z}^2.$$

Theorem (Montgomery, 1988)**.** *Among all lattices $\Lambda\subset\mathbb{R}^2$ with fixed density,*

$$\max_{z\in\mathbb{R}^2} E_\Lambda(z;\alpha) \qquad \textit{is minimized}$$

if and only if Λ is the hexagonal lattice Λ_2.

Main Result (Bétermin, Faulhuber, Steinerberger, 2021)**.** *Among all lattices $\Lambda\subset\mathbb{R}^2$ with fixed density,*

$$\min_{z\in\mathbb{R}^2} E_\Lambda(z;\alpha) \qquad \textit{is maximized}$$

if and only if Λ is the hexagonal lattice Λ_2.

One nice aspect of Montgomery's result is that the maximum is assumed in a lattice point; in contrast, we have relatively little control over the point z in which the minimum is assumed (see Figure 1) which makes the proof significantly harder. One important consequence of our Main Result is that the hexagonal lattice maximizes the minimum while *simultaneously* minimizing the maximum of E_Λ (the latter being due to Montgomery). We expect this to be a very rare property if (1) is generalized to higher dimensions. This reaffirms the special role that the hexagonal lattice Λ_2 plays for variational problems in $\mathbb{R}^2$. The Main Result has many consequences, one of which is, in combination with Montgomery's result, that the conjecture of Strohmer and Beaver [9] is finally affirmed.

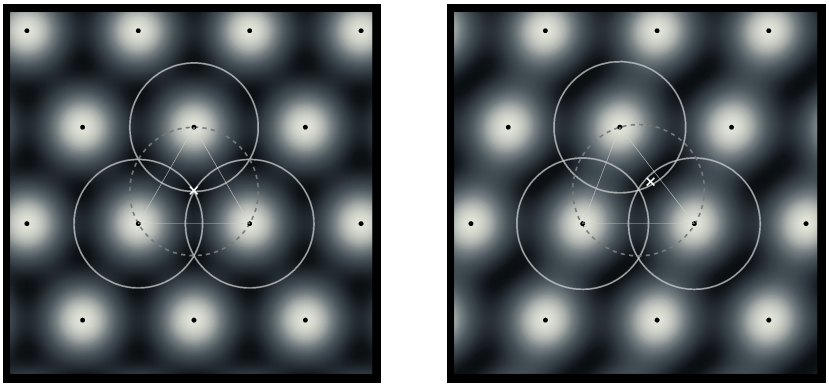


FIGURE 1. Gaussian lattice sum for the hexagonal lattice Λ_2 (left) and a non-hexagonal, non-rectangular lattice (right). For $E_{\Lambda_2}(z;\alpha)$ the minimum among all z is attained at the circumcenter for all $\alpha > 0$ and the closest lattice points are a covering radius away. For general lattices, the minimum of $E_\Lambda(z;\alpha)$, marked by $\times$, is "close" to the circumcenter and varies with α (see also [1]).

2. Consequences for Gaussian Gabor frames

Consider the Gaussian Gabor system $\mathcal{G}(\varphi,\Lambda)$, $\Lambda = M\mathbb{Z}^2$ with $M \in GL(2,\mathbb{R})$ and

$$\mathrm{vol}(\Lambda) = |\det(M)| \quad \text{is fixed.}$$

The quantity $\delta(\Lambda) = \mathrm{vol}(\Lambda)^{-1}$ is called the density of the lattice and determines the (over-)sampling rate of the Gabor system $\mathcal{G}(\varphi,\Lambda)$. The window function

$$\varphi(t) = 2^{1/4} e^{-\pi t^2}$$

is the Fourier invariant standard Gaussian of $L^2(\mathbb{R})$-unit norm. The associated frame operator $S_\Lambda : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ is given by

$$S_\Lambda f = \sum_{\lambda\in\Lambda} \langle f, \pi(\lambda)\varphi\rangle\, \pi(\lambda)\varphi.$$

Here, $\pi(\lambda)$ denotes a shift in the time-frequency plane by $\lambda \in \mathbb{R}^2$;

$$\pi(\lambda)\varphi(t) = M_\omega T_x \varphi(t) = \varphi(t-x)e^{2\pi i\omega t}, \quad \lambda = (x,\omega).$$

The Gabor system is a frame if and only if the frame inequality is fulfilled;

$$A_\Lambda \|f\|_{L^2}^2 \le \sum_{\lambda\in\Lambda} |\langle f, \pi(\lambda)\varphi\rangle|^2 \le B_\Lambda \|f\|_{L^2}^2, \quad \forall f \in L^2(\mathbb{R}),$$

for positive constants $0 < A_\Lambda \le B_\Lambda < \infty$. The sharpest possible constants above are obtained from the spectral bounds of the frame operator S_Λ;

$$A_\Lambda^{-1} = \left\|S_\Lambda^{-1}\right\|_{L^2\to L^2} \quad \text{and} \quad B_\Lambda = \|S_\Lambda\|_{L^2\to L^2}.$$

Due to the results of Lyubarskii [5], Seip [7], and Seip and Wallsten [8], we know that $\mathcal{G}(\varphi, \Lambda)$ is a frame whenever $\delta(\Lambda) > 1$. Hence, it makes sense to ask for the optimal sampling pattern for a fixed oversampling rate δ.

In [9], Strohmer and Beaver compared the frame condition number, i.e., B_Λ/A_Λ, of Gaussian Gabor systems with square lattice and hexagonal lattice of density 2. They observed that, in this case, the condition number for the square lattice is $\sqrt{2}$ and for the hexagonal lattice it is "*suspiciously close*" to $\sqrt[3]{2}$. Furthermore, they conjectured that this is the smallest possible condition number among all lattices of density 2. Further numerical comparisons also suggest that the hexagonal lattice gives the optimal condition number for any fixed oversampling rate.

In general, it is extremely hard to compute frame bounds of Gabor systems, but for Gaussian Gabor systems of oversampling rate $2n$, $n \in \mathbb{N}$, a result of Janssen [4] shows that the computation of the sharp frame bound reduces to finding the minimum and maximum of a Gaussian lattice sum as given by (1). By using Montgomery's result [6], it was shown in [3] that in this case B_Λ is minimized only for the hexagonal lattice. Using the Main Result [2], we also see that A_Λ is maximized in this case. Therefore, the conjecture of Strohmer and Beaver is also confirmed in this case. We expect that in higher dimensions the optimal lattices for the lower and upper frame bound will in general differ from each other.

The Main Result has further remarkable consequences, which are discussed in detail in [2]. Due to the universality of Gaussian lattice sums and its wide applicability in different areas of mathematics, it stands to reason that the Main Result also has many more implications, not just the ones stated in [2].

References

[1] Albert Baernstein II. *A minimum problem for heat kernels of flat tori*, in Extremal Riemann surfaces vol. 201 of Contemporary Mathematics, AMS (1997), 227–243.

[2] L. Bétermin, M. Faulhuber and S. Steinerberger. *A variational principle for Gaussian lattice sums*, arXiv preprint: 2110.06008 (2021), 1–62.

[3] Markus Faulhuber. *Minimal Frame Operator Norms Via Minimal Theta Functions*, Journal of Fourier Analysis and Applications, **24(2)** (2018), 545–559.

[4] Augustus J. E. M. Janssen. *Some Weyl-Heisenberg frame bound calculations*, Indagationes Mathematicae, **7(2)** (1996), 165–183.

[5] Yurii Lyubarskii. *Frames in the Bargmann space of entire functions*, in Entire and Subharmonic Functions vol. 11 of Advances in Soviet Mathematics, AMS (1992), 167–180.

[6] Hugh L. Montgomery, *Minimal theta functions*, Glasgow Mathematical Journal, **30(1)** (1988), 75–85.

[7] Kristian Seip. *Density theorems for sampling and interpolation in the Bargmann–Fock space I.* Journal für die reine und angewandte Mathematik (Crelles Journal), **429** (1992), 91–106.

[8] Kristian Seip and Robert Wallstén. *Density theorems for sampling and interpolation in the Bargmann–Fock space II*, Journal für die reine und angewandte Mathematik (Crelles Journal), **429** (1992), 107–114.

[9] T. Strohmer and S. Beaver. *Optimal OFDM design for time-frequency dispersive channels*, Communications, IEEE Transactions, **51(7)** (2003), 1111–1122.

Using matrix factorizations for interpretability

Deanna Needell
(joint work with Laura Balzano, Hanbaek Lyu)

In modern data analysis, a central step is to find a low-dimensional representation to better understand, compress, or convey the key phenomena captured in the data. Matrix factorization provides a powerful setting for one to describe data in terms of a linear combination of factors or atoms. In this setting, we have a data matrix $X \in \mathbb{R}^{d\times n}$, and we seek a factorization of X into the product WH for $W \in \mathbb{R}^{d\times r}$ and $H \in \mathbb{R}^{r\times n}$. This problem has gone by many names over the decades, each with different constraints: dictionary learning, factor analysis, topic modeling, component analysis. It has applications in text analysis, image reconstruction, medical imaging, bioinformatics, and many other scientific fields more generally [2–5, 9, 10, 12].

Online matrix factorization is a problem setting where data are accessed in a streaming manner and the matrix factors should be updated each time. That is, we get draws of X from some distribution π and seek the best factorization such that the expected loss $\mathbb{E}_{X\sim\pi}\left[\|X - WH\|_F^2\right]$ is small. This is a relevant setting in today's data world, where large companies, scientific instruments, and healthcare systems are collecting massive amounts of data every day. One cannot compute with the entire dataset, and so we must develop online algorithms to perform the computation of interest while accessing them sequentially.

A natural way to relax the assumption of independence in this online context is through the Markovian assumption. In many cases one may assume that the data are not independent, but independent conditioned on the previous iteration. The central contribution of our work is to extend the analysis of online matrix factorization in [8] to the setting where the sequential data form a Markov chain. This is naturally motivated by the fact that the Markov chain Monte Carlo (MCMC) method is one of the most versatile sampling techniques across many disciplines, where one designs a Markov chain exploring the sample space that converges to the target distribution.

In this paper, we analyze convergence properties of the following scheme of OMF:

(1)

$$\textit{Upon arrival of } X_t\textit{:} \quad \begin{cases} H_t = \arg\min_{H\in\mathcal{C}'\subseteq\mathbb{R}^{r\times n}} \|X_t - W_{t-1}H\|_F^2 + \lambda\|H\|_1 \\ A_t = (1-w_t)A_{t-1} + w_t H_t H_t^T \\ B_t = (1-w_t)B_{t-1} + w_t H_t X_t^T \\ W_t = \arg\min_{W\in\mathcal{C}\subseteq\mathbb{R}^{d\times r}} \left(\mathrm{tr}(WA_tW^T) - 2\mathrm{tr}(WB_t)\right) \\ \qquad \text{s.t. } \mathrm{tr}((B_t^T - WA_t)(W_{t-1} - W)^T) \le 0 \end{cases}$$

where $(w_t)_{t\ge1}$ is a prescribed sequence of weights, and A_0 and B_0 are zero matrices of size $r \times r$ and $r \times d$, respectively. Note that the L_2-loss function is augmented with the L_1-regularization term $\lambda\|H\|_1$ with regularization parameter $\lambda \ge 0$, which forces the code H_t to be sparse.

This result is summarized in Theorem 1, rigorously establisheing convergence of a non-convex generalization (1) of the online matrix factorization scheme in [7, 8] when the data sequence $(X_t)_{t\geq 0}$ is realized as a function of some underlying Markov chain with a mild mixing condition.

Theorem 1. *We assume mild conditions on the Markov chain and loss functions (see [6]). Let $(W_t, H_t)_{t\geq 1}$ be a solution to the optimization problem* (1)*. Then the following hold.*

(i): $\lim_{t\to\infty} \mathbb{E}[f_t(W_t)] = \lim_{t\to\infty} \mathbb{E}[\hat{f}_t(W_t)] < \infty$.

(ii): $f_t(W_t) - \hat{f}_t(W_t) \to 0$ *and* $f(W_t) - \hat{f}_t(W_t) \to 0$ *as* $t \to \infty$ *almost surely.*

(iii): *Almost surely,*

$$\lim_{t\to\infty} \|\nabla_W f(W_t) - 2(W_t A_t - B_t)\|_F = 0. \tag{2}$$

Furthermore, the distance between W_t and the set of all local extrema of f in $\mathcal{C}$ converges to zero almost surely.

We remark here that the algorithm and theory have also been extended to the tensor case, see [11] for details.

We now present an example of how one can use ONMF and its tensor version in applications. We demonstrate our method on video data of brain activity across a mouse cortex, and how our online method learns dictionaries for the spatial and temporal activation patterns simultaneously. The original video is due to Barson et al. by using genetically encoded calcium indicators to image brain activity transcranially [1].

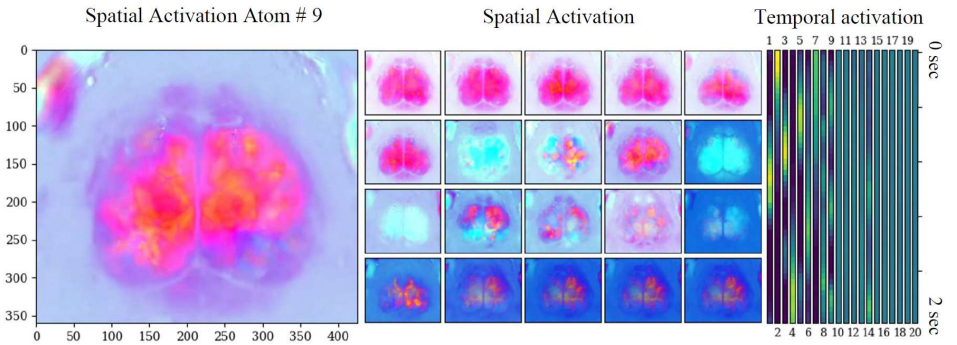


FIGURE 1. Learning 20 CP-dictionary patches from video frames on brain activity across the mouse cortex.

Our algorithm learns a dictionary in the space-color mode that shows spatial activation patterns and the corresponding time mode shows their temporal activation pattern, as seen in Figure 1. Due to the nonnegativity constraint, spatial activation atoms representing localized activation regions in the cortex are learned, while the darker ones represent the background brain shape without activation. On the other hand, the activation frequency is simultaneously learned by the

temporal activation atoms shown in Figure 1 (right). For instance, the spacial activation atom # 9 (numbered lexicographically) activates three times in its corresponding temporal activation atom in the right, so such activation pattern has an approximate frequency of 2/3 sec.

References

[1] Daniel Barson, Ali S Hamodi, Xilin Shen, Gyorgy Lur, R Todd Constable, Jessica A Cardin, Michael C Crair, and Michael J Higley. Simultaneous mesoscopic and two-photon imaging of neuronal activity in cortical circuits. *Nature methods*, 17(1):107–113, 2020.

[2] Michael W Berry and Murray Browne. Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, 11(3):249–264, 2005.

[3] Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173, 2007.

[4] Rostyslav Boutchko, Debasis Mitra, Suzanne L Baker, William J Jagust, and Grant T Gullberg. Clustering-initiated factor analysis application for tissue classification in dynamic brain positron emission tomography. *Journal of Cerebral Blood Flow & Metabolism*, 35(7):1104–1111, 2015.

[5] Yang Chen, Xiao Wang, Cong Shi, Eng Keong Lua, Xiaoming Fu, Beixing Deng, and Xing Li. Phoenix: A weight-based network coordinate system using matrix factorization. *IEEE Transactions on Network and Service Management*, 8(4):334–347, 2011.

[6] Hanbaek Lyu, Deanna Needell, and Laura Balzano. Online matrix factorization for Markovian data and applications to network dictionary learning. *arXiv:1911.01931*, 2019.

[7] Julien Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, pages 2283–2291, 2013.

[8] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[9] Bin Ren, Laurent Pueyo, Guangtun Ben Zhu, John Debes, and Gaspard Duchêne. Nonnegative matrix factorization: robust extraction of extended structures. *The Astrophysical Journal*, 852(2):104, 2018.

[10] Arkadiusz Sitek, Grant T Gullberg, and Ronald H Huesman. Correction for ambiguous solutions in factor analysis using a penalized least squares objective. *IEEE transactions on medical imaging*, 21(3):216–225, 2002.

[11] C. Strohmeier, H. Lyu, and D. Needell. Online nonnegative tensor factorization and cp-dictionary learning for markovian data. *NeurIPS Opt+ML Workshop*, 2020.

[12] Leo Taslaman and Björn Nilsson. A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. *PloS One*, 7(11):e46331, 2012.

Machine learning meets super-resolution

Hrushikesh N. Mhaskar

In this talk, we pointed out a duality between the question of approximation of functions and the question of blind source signal separation. We also explained our idea that treating the classification problem as a super-resolution problem leads to a provably good classification using a small number of samples, overcoming the problem of non-disjoint class boundaries.

The question of blind source signal separation can be formulated in a general setting as follows. Let $\mathbb{X}$ be a locally compact, metric, measure space,

$\mu = \sum_{k=1}^{K} a_k \delta_{w_k}$ for some integer $K \geq 2$, $a_k \in \mathbb{R}$, $w_k \in \mathbb{X}$, where δ_{w_k} is the Dirac delta distribution supported at w_k. With a kernel $G : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, we observe

$$f(y_j) = \sum_{k=1}^{K} a_k G(y_j, w_k) + \epsilon_j, \qquad j = 1, \cdots, M, \tag{1}$$

for judiciously chosen $y_j \in \mathbb{X}$, where ϵ_j are realizations of a mean zero random variable with an unknown distribution. The problem is then to recuparate K, a_k, w_k, and estimate the accuracy in terms of M.

Nominally, the problem of kernel based approximation of f is also the same, except that f is considered to be an unknown function, one has typically no control on the choice of y_j, and the parameters as K, a_k, w_k may be chosen judiciously to achieve a good approximation, the approximation error being estimated in terms of M.

Table 3 gives a brief comparison.

Machine learning	Signal separation
Approximate f, **Choose** a_k, w_k	f *is given*, **find** a_k, w_k
Error measured in *function space norm*	Error measured by *numerical accuracy*
M depends upon **smoothness (prior) on** f	M depends upon **minimal separation**
Solve regularization problem	Solve regularization problem

TABLE 3. A formal comparison between kernel based approximation of functions and blind source signal separation.

A crucial ingredient in our solution to both the problems is a family of localized kernels $\Phi_n \in C(\mathbb{X} \times \mathbb{X})$ satisfying

$$|\Phi_n(y, z)| \leq c \frac{n^q}{\max(1, (n\rho(y, z))^S)}, \qquad n \geq 1, \ y, z \in \mathbb{X}, \tag{2}$$

where ρ is the metric on $\mathbb{X}$, q serves as the "dimension" of $\mathbb{X}$, $S > q$ is a parameter depending on the construction of the kernels, and $c > 0$ is a constant depending upon $\mathbb{X}$ and S, but independent of n, y, and z. For example, if $\mathbb{X}$ is a q-dimensional compact, smooth, orientable manifold, λ_k^2's are the eigenvalues of the (negative) Laplace-Beltrami operator, and ϕ_k's are the corresponding eigenfunctions, then the family Φ_n can be constructed by the formula

$$\Phi_n(y, z) = \sum_{\ell=0}^{\infty} H\left(\frac{\lambda_\ell}{n}\right) \phi_\ell(y)\phi_\ell(z), \tag{3}$$

where $H : [0, \infty) \to [0, 1]$ is an infinitely differentiable function, equal to 1 on $[0, 1/2]$, equal to 0 on $[1, \infty)$, and non-increasing on $[1/2, 1]$ [6,8]. We have proved

in [5,8] that if μ^* is a positive measure on $\mathbb{X}$, then under some additional conditions,

$$\sigma_n(f) := \int_{\mathbb{X}} f(z)\Phi_n(\circ, z)d\mu^*(z) \to f \tag{4}$$

uniformly for all $f \in C(\mathbb{X})$, and estimated the degree of approximation. In view of this fact, it is clear that for any (possibly signed) measure μ having bounded total variation on $\mathbb{X}$ (i.e., $\mu \in C(\mathbb{X})^*$),

$$\sigma_n(\mu) := \left\{ \int_{\mathbb{X}} \Phi_n(\circ, z)d\mu(z) \right\} d\mu^* \to \mu \tag{5}$$

in the weak-* sense. In [7], we have introduced a discrepancy expression to measure the rate of weak-* convergence. Our discrepancy is a generalization of the Erdős-Turán discrepancy as well as an analogue of the Wasserstein distance. We have discussed in detail the rate of convergence using this discrepancy, and proved in particular, that in the context of blind source signal separation ($\mu = \sum_{k=1}^K a_k\delta_{w_k}$), the minimum amount of information M needed to achieve signal separation in a robust manner is $= \Omega(\eta^{-q/\beta})$, where η is the minimal separation among the points w_k, and β is a parameter that defines the discrepancy. The term super-resolution is used to consider the question of how to overcome this barrier.

When $\mu = \sum_{k=1}^K a_k\delta_{w_k}$, the density of $\sigma_n(\mu)$ is $\sum_{k=1}^K a_k\Phi_n(\circ, w_k)$. Since Φ_n is highly localized, the peaks of this density occur approximately at the points w_k's and the corresponding value approximates a_k. Moreover, for a properly chosen threshold θ, the region where the absolute value of this density is larger than θ splits into exactly K clusters, separated by half the minimal separation among the w_k's. Thus, we obtain an algorithm for finding all the parameters, including the number of clusters K. This is illustrated in many of our papers, e.g., [1–3, 9]. We note that for (4) to hold with the correct rate of convergence without saturation, it is important that Φ_n should **not be a positive kernel**. For the convergence of $\sigma_n(\mu)$ to μ when μ is finitely supported, the localization of the kernels avoids interference from one point w_k to another. When μ is not finitely supported, then the correct rate of approximation requires that Φ_n should not positive, but this property makes it impossible to avoid interference so that the negative part of the kernel at some point in the support of μ is not canceled by the positive part of the kernel from another part of the support of μ. Thus, the question of approximation of μ itself needs to be treated separately from the question of finding the support of μ. This is accomplished in the same manner using Φ_n^2 in place of Φ_n.

The second major insight discussed in our talk is to consider the problem of classification as the problem of super-resolution. The problem of classification is the following. We have data of the form $\{(y_j, z_j)\}_{j=1}^M$ where y_j's are samples from a probability distribution μ, and z_j is the class label corresponding to y_j; i.e., an element of $\{1, \cdots, K\}$. The question is to predict the class label for a new point y sampled from μ. In active learning paradigm, all the points y_j are known, but none of the labels z_j to start with. The objective is to choose the points y_j judiciously at which the corresponding label ought to be queried, and then extend these labels to the rest of the support of μ. Considering the conditional probabilities of y

belonging to different classes, we obtain $\mu = \sum_{k=1}^{K} \mu_k$, where μ_k's are positive measures, with the support $\mathcal{S}_k$ of each μ_k being the set of all y's with class label k. Thus, if we could determine the support $\mathcal{S}_k$ for each k, then the classification problem is solved using exactly one y_k from $\mathcal{S}_k$; the minimal number of samples required to solve the problem. If each μ_k has the form $a_k\delta_{w_k}$, then this is exactly the problem of signal separation. However, unlike in that problem the minimal separation among $\mathcal{S}_k$'s is typically 0. Therefore, this is a super-resolution problem, except that we are interested only in separating $\mathcal{S}_k$ rather than approximating the μ_k's themselves. As such, we have proposed in [4] to use the same method as for signal separation as in our previous papers, but with Φ_n^2 in place of Φ_n. In [4], the kernel Φ_n is constructed using Hermite polynomials, with localization proved using the Mehler identity and the Tauberian theorem in [6]. We have proved that the resulting classification scheme converges in the sense of F-score. We note that the number of classes is an output of our theorem; we do not need to know a priori how many classes are in the data. Indeed, we have argued in [4] that the labels can be hierarchical in nature.

References

[1] C. K. Chui and H. N. Mhaskar. *Signal decomposition and analysis via extraction of frequencies*, Applied and Computational Harmonic Analysis, **40**(1):97–136, 2016.

[2] C. K. Chui and H. N. Mhaskar, *A unified method for super-resolution recovery and real exponential-sum separation*, Appl. Comput. Harmon. Anal., **46**(2):431–451, March 2019.

[3] C. K. Chui, H. N. Mhaskar, and M. D. van der Walt, *Data-driven atomic decomposition via frequency extraction of intrinsic mode functions*, GEM-International Journal on Geomathematics, **7**(1):117–146, 2016.

[4] A. Cloninger and H. Mhaskar, *Cautious active clustering*, Applied and Computational Harmonic Analysis, **54**:44–74, 2021.

[5] M. Maggioni and H. N. Mhaskar, *Diffusion polynomial frames on metric measure spaces*, Applied and Computational Harmonic Analysis, **24**(3):329–353, 2008.

[6] H. N. Mhaskar, *A unified framework for harmonic analysis of functions on directed graphs and changing data*, Appl. Comput. Harm. Anal., **44**(3):611–644, 2018.

[7] H. N. Mhaskar, *Super-resolution meets machine learning: approximation of measures*, Journal of Fourier Analysis and Applications, **25**(6):3104–3122, 2019.

[8] H. N. Mhaskar, *Kernel-based analysis of massive data*, Frontiers in Applied Mathematics and Statistics, **6**:30, 2020.

[9] H. N. Mhaskar and J. Prestin, *On local smoothness classes of periodic functions*, Journal of Fourier Analysis and Applications, **11**(3):353–373, 2005.

Solving Inverse Problems With Deep Neural Networks – Robustness Included?

Martin Genzel

(joint work with Jan Macdonald and Maximilian März)

In recent years, deep learning methods have been successfully applied to many problems of the natural sciences [7]. A prominent example of such scientific machine learning is the development of efficient solutions strategies for *inverse problems*, such as those encountered in medical imaging (see Fig. 1 for an example).

Formally, an inverse problem in its prototypical, finite-dimensional form reads as follows:

$$\text{(IP)} \quad \left\{ \begin{array}{c} \textit{Given a forward operator } \mathcal{A}\colon \mathbb{R}^d \to \mathbb{R}^m \textit{ and corrupted measurements} \\ \boldsymbol{y} = \mathcal{A}(\boldsymbol{x}_0) + \boldsymbol{e} \textit{ with } \|\boldsymbol{e}\|_2 \leq \eta, \textit{ reconstruct the signal } \boldsymbol{x}_0 \in \mathbb{R}^d. \end{array} \right\}$$

Here, $\boldsymbol{e} \in \mathbb{R}^m$ models an unknown *perturbation* of the measurement process, e.g., caused by noise or misspecifications in the forward operator.

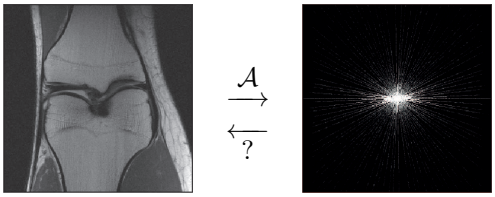


FIGURE 1. Example from the NYU fastMRI dataset for magnetic resonance imaging (MRI) [6]. Here, the inverse problem (IP) basically amounts to reconstructing an image signal (left) from highly undersampled frequency measurements (right). Thus, $\mathcal{A}$ is the composition of a Fourier transform with a binary sampling mask.

Until 2016, the gold standard for inverse problems was given by variational methods, typically phrased as (convex) optimization schemes with handcrafted regularization terms [3, 4]. Since then, *deep-learning-based solutions* have revolutionized the field, often clearly outperforming classical algorithms in terms of precision and speed [2, 8]. Many of these approaches rely on a *supervised* learning procedure, which in its most basic form looks as follows: Given a large set of example pairs $\{(\boldsymbol{y}^i, \boldsymbol{x}_0^i)\}_{i=1}^M$ drawn from (IP), one intends to compute a reconstruction neural network $\mathrm{Net}[\hat{\boldsymbol{\theta}}]\colon \mathbb{R}^m \to \mathbb{R}^d$ by means of empirical risk minimization, i.e., the network weights $\hat{\boldsymbol{\theta}}$ are an (approximate) solution to

$$\text{(ERM)} \qquad \min_{\boldsymbol{\theta}} \ \tfrac{1}{M} \textstyle\sum_{i=1}^M \ell\left(\mathrm{Net}[\boldsymbol{\theta}](\boldsymbol{y}^i), \boldsymbol{x}_0^i\right),$$

where $\ell\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is an appropriate loss function. The hope is that the resulting map does not only fit the training data but also generalizes well to unseen instances $(\boldsymbol{y}, \boldsymbol{x}_0)$ of (IP) in the sense that $\mathrm{Net}[\hat{\boldsymbol{\theta}}](\boldsymbol{y}) \approx \boldsymbol{x}_0$.

The *robustness* against noisy perturbations is arguably one of the most important features of an inversion method (no matter if learned or not). Formally, we call a solution map $\mathrm{Rec}\colon \mathbb{R}^m \to \mathbb{R}^d$ for (IP) *robust* with respect to the input $\boldsymbol{y} = \mathcal{A}(\boldsymbol{x}_0) + \boldsymbol{e}$ if it satisfies

$$\|\mathrm{Rec}\big(\mathcal{A}(\boldsymbol{x}_0)\big) - \mathrm{Rec}(\boldsymbol{y})\|_2 \leq C \cdot \eta \tag{1}$$

for a small constant $C > 0$. In other words, the error due to the perturbation $\boldsymbol{e}$ is controlled through the admissible noise level η. A valid point in favor of model-based variational methods is that bounds of the form (1) can be often theoretically verified under appropriate assumptions on $\mathcal{A}$ and the regularizer, e.g., see [4].

In contrast, there are currently no comprehensive guarantees for practical data-driven schemes available. Hence, an empirical verification of their robustness remains indispensable. While a remarkable resilience against statistical noise has been frequently reported, several alarming findings indicate that inversion procedures via deep neural networks might become unstable [1]. In analogy to *adversarial attacks* on classifiers [9], it was pointed out that even tiny distortions in the input domain may cause severe artifacts in the recovered (image) signals; thus, unlike (1), the reconstruction $\mathrm{Rec}(\boldsymbol{y})$ would drastically deviate from its unperturbed counterpart $\mathrm{Rec}\big(\mathcal{A}(\boldsymbol{x}_0)\big)$. If this suspicion is substantiated, it would certainly have detrimental implications on the usage of artificial intelligence in safety-critical fields such as medical imaging.

Addressing this controversial debate, the presented work [5] is the first to draw a fairly different, far more optimistic, picture of the aforementioned issues. In our large-scale study, we have investigated the reliability of deep learning algorithms for solving inverse problems. Our experimental designs cover standard compressed sensing as well as image recovery from Fourier and Radon measurements, including a real-world scenario from medical imaging based on the NYU fastMRI dataset [6] (see Fig. 1). In view of previous concerns, the outcome of our research was quite unexpected: standard end-to-end neural network models can not only resist against statistical noise, but also against *adversarial perturbations*. The latter amounts to computing a worst-case perturbation of the (noiseless) measurements $\boldsymbol{y}_0 := \mathcal{A}(\boldsymbol{x}_0)$ such that the error of a given solver $\mathrm{Rec}\colon \mathbb{R}^m \to \mathbb{R}^d$ for (IP) is maximized:

$$\boldsymbol{e}_{\mathrm{adv}} \in \mathrm{argmax}_{\|\boldsymbol{e}\|_2 \leq \eta} \, \|\mathrm{Rec}(\boldsymbol{y}_0 + \boldsymbol{e}) - \boldsymbol{x}_0\|_2 \, . \tag{2}$$

In this way, we were able to demonstrate empirically that learning-based schemes indeed have the potential to obey error bounds of form (1). A distinctive feature of our approach is the quantitative and qualitative comparison with total-variation minimization, which serves as a provably robust reference method (see Fig. 2 for two example results from [5]). On the other hand, we have also identified pathological situations where instabilities (error blowups) are possible. Importantly, all considered neural networks were obtained by ordinary empirical risk minimization (ERM), without the need for an adversarial defense.

References

[1] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, *On instabilities of deep learning in image reconstruction and the potential costs of AI*, Proc. Natl. Acad. Sci. **117** (2020), no. 48, 30088–30095.

[2] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, *Solving inverse problems using data-driven models*, Acta Numer. **28** (2019), 1–174.

[3] M. Benning and M. Burger, *Modern regularization methods for inverse problems*, Acta Numer. **27** (2018), 1–111.

[4] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, Birkhäuser, 2013.

[5] M. Genzel, J. Macdonald, and M. März, *Solving inverse problems with deep neural networks – Robustness included?* arXiv:2011.04268, 2020.

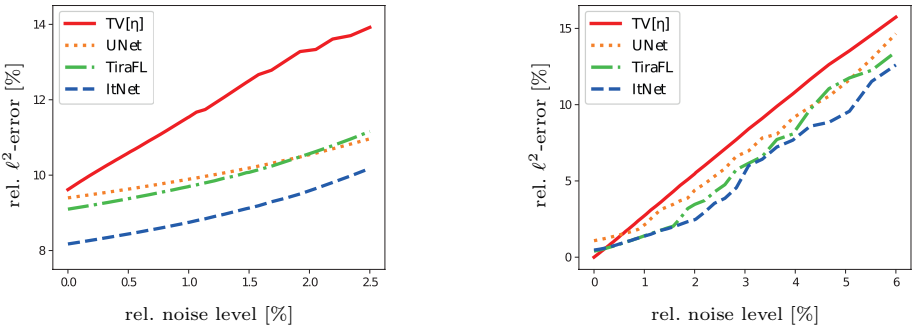


FIGURE 2. The left plot shows the impact of *adversarial perturbations* on the MRI signal from Fig. 1. The recovery error of total-variation minimization TV[η] is compared to three neural-network-based solvers (UNet, TiraFL, and ItNet). The corresponding error curves are generated by computing $\boldsymbol{e}_{\text{adv}}$ according to (2) for a range of noise levels η (and every method independently). The right plot shows an analogous experiment in the idealistic situation of one-dimensional piecewise constant signals. Here, the benchmark TV[η] is a perfect match that allows for exact recovery from noiseless measurements. Images taken from [5].

[6] F. Knoll, J. Zbontar, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdzalv, A. Romero, M. Rabbat, P. Vincent, J. Pinkerton, D. Wang, N. Yakubova, E. Owens, C. L. Zitnick, M. P. Recht, D. K. Sodickson, and Y. W. Lui, *fastMRI: a publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning*, Radiology Artif. Intell. **2** (2020), no. 1, e190007.
[7] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature **521** (2015), no. 7553, 436–444.
[8] G. Ongie, A. Jalal, R. G. Baraniuk, C. A. Metzler, A. G. Dimakis, and R. Willett, *Deep learning techniques for inverse problems in imaging*, IEEE J. Sel. Areas Inf. Theory **1** (2020), no. 1, 39–56.
[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, *Intriguing properties of neural networks*, arXiv:1312.6199, 2013.

Intersectionless Envelope Estimation for EMD

LASLO HUNHOLD

The Empirical Mode Decomposition (EMD) is a heuristic self-adaptive and data-driven method for additively separating a multi-component, nonlinear, nonstationary signal $s\colon [0,1] \to \mathbb{R}$ into Intrinsic Mode Functions (IMFs) and a residual (see [1]). IMFs have the form

$$u(t) = a(t) \cdot \cos(\phi(t)) \tag{1}$$

with 'amplitude' $a(t) > 0$ and 'phase' $\phi(t)$. The 'frequency' $\phi'(t) > 0$ and $a(t)$ are supposed to be 'slowly varying'. In this abstract we will only consider a single

EMD iteration, namely separating a given signal $s(t)$ into an IMF $u(t)$ and a residual $r(t)$:

$$s(t) = u(t) + r(t). \tag{2}$$

Subsequent EMD iterations are carried out by respectively taking the residual of the previous iteration as the input signal, yielding the desired additive separation of a signal into multiple IMFs and a single residual (that of the last iteration).

The separation process in (2) is called 'sifting' and illustrated in Figure 1: The first step is estimating the lower envelope $m_-(t)$ and upper envelope $m_+(t)$, where an envelope is a curve that traces the signal's extremes without intersecting it. The residual $r(t)$ and IMF $u(t)$ are then obtained using

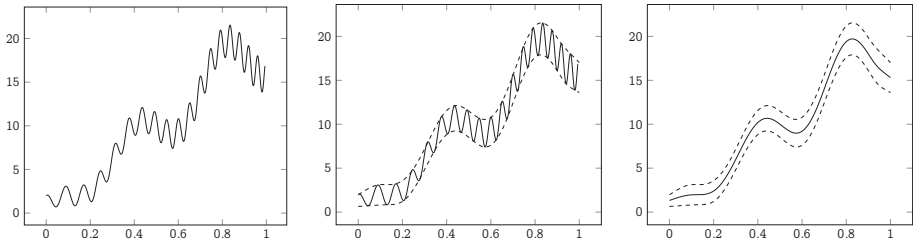


FIGURE 1. Given a signal $s(t)$ (left) its envelopes (dashed) are estimated (middle) and averaged, yielding the residual $r(t)$ (right).

$$r(t) = \frac{m_-(t) + m_+(t)}{2}, \tag{3}$$

$$u(t) = s(t) - r(t). \tag{4}$$

In particular, we can directly determine $a(t) = m_+(t) - r(t)$ and normalize the IMF to $u(t)/a(t) = \cos(\phi(t))$, leaving only the determination of $\phi(t)$. This phase extraction will however not be in the scope of this abstract and we want to focus on the envelope estimation.

Without loss of generality we only consider the estimation of the upper envelope. We can estimate the lower envelope by negating the estimated upper envelope of the negated signal. In the classical approach (see [1] and [2]), the upper envelope is estimated by natural cubic B-Spline interpolation of the signal's local maxima. This has the disadvantage that the estimated envelope can intersect with the signal, violating the envelope property (see Figure 2). There have been numerous approaches to improve the mathematical formulation and estimation of envelopes (see for example [3] and [4]) with their own downsides in regard to precision, efficiency and generalization for multivariate data. EMD is widely used (see for example [5]), further underlining the need for a better envelope estimation method.

The author proposes the following iterative scheme to obtain better envelopes: For an envelope estimate $\tilde{m}$ (that is initially set to 0), determine the spots t where $s'(t) = \tilde{m}'(t)$ and $s''(t) < \tilde{m}''(t)$ or $t \in \{0, 1\}$. Natural cubic B-Spline interpolation

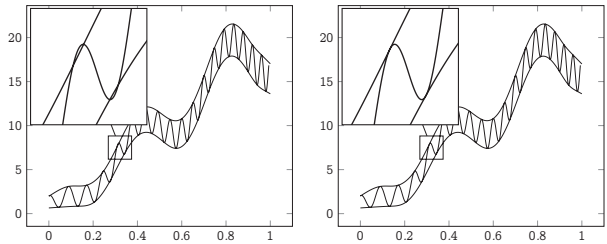


FIGURE 2. The classical method (left) yields intersecting envelopes, whereas the proposed iterative method (right) prevents intersections.

of these spots yields a new estimate m that can be used as $\tilde{m}$ for the next iteration. This is repeated until m is at most ε below s at any given point (see Algorithm 1 for the multivariate case with Hessian $\boldsymbol{H}$). Specifically, the first iteration is identical to the classical method of interpolating local maxima.

Algorithm 1 Iterative Slope Envelope Estimation for multivariate signals

input : multicomponent signal $s \in \mathcal{C}^2([0,1]^d, \mathbb{R})$
tolerance $\varepsilon > 0$
output: upper envelope $m \in \mathcal{C}^2([0,1]^d, \mathbb{R})$

$m \longleftarrow 0 \in \mathcal{C}^0([0,1]^d, \mathbb{R})$;
repeat
  $\tilde{m} \longleftarrow m$;
  $P \longleftarrow \{(\boldsymbol{t}, s(\boldsymbol{t})) \in [0,1]^d \times \mathbb{R} \mid \nabla(s-\tilde{m})(\boldsymbol{t}) = \boldsymbol{0} \wedge \boldsymbol{H}(s-\tilde{m})(\boldsymbol{t})$ negative definite$\}$;
  $m \longleftarrow$ Interpolate$(P \cup (\partial[0,1]^d, s(\partial[0,1]^d)))$;
until $\|\max(s-m, 0)\|_\infty < \varepsilon$;

The algorithm shows no asymptotic overhead compared to the classical method: Finding local maxima and evaluating $\|\max(s-m,0)\|_\infty$ are embarassingly parallel (i.e. trivially separable into parallel tasks) and natural cubic B-Spline interpolation is linearly complex. The relatively small number of necessary iterations can at least be confirmed empirically (see Figure 3).

Theoretically speaking, the fast convergence can be explained with the fact that the interpolation points change less and less with each iteration. The effect of such changes on the slope in other areas is additionally dampened by the interpolation. This indicates a self-stabilizing behaviour whose proof would require an estimation of the effect of control point changes on the interpolation's slope.

All in all, the iterative slope envelope estimation yields intersectionless envelopes with negligible overhead. Within the EMD heuristic, this may improve the quality of the signal separation.

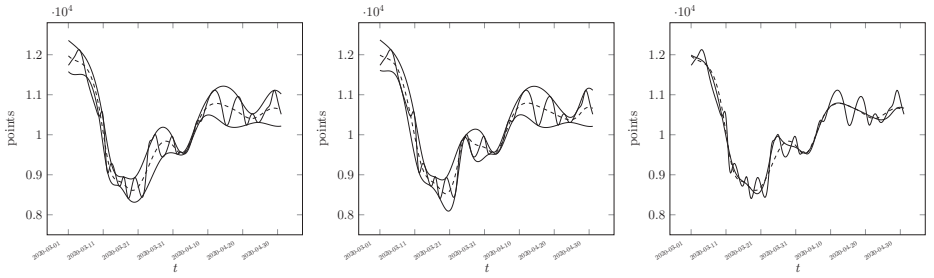


FIGURE 3. Subsection from March 1st 2020 to May 1st 2020 of daily DAX closing prices (thin). The algorithm converged on the range January 2nd 2020 to November 25th 2021 after 4 iterations with $\varepsilon = 10$. Left and middle show estimated envelopes (thick) and residuals (dashed). The classical method (left) exhibits the intersection problem, whereas the iterative slope method (middle) yields intersectionless envelopes. The residuals (right) of the classical (dashed) and iterative method (thick) differ significantly.

REFERENCES

[1] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung and Henry H. Liu. 'The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis'. In: *Proceedings of the Royal Society A*. Vol. 454. 1971. Mar. 1998, pp. 903–995. DOI: 10.1098/rspa.1998.0193.

[2] Quihui Chen, Norden E. Huang, Sherman Riemenschneider and Yuesheng Xu. 'A B-Spline Approach for Empirical Mode Decompositions'. In: *Advances in Computational Mathematics* 24 (Jan. 2006), pp. 171–195. DOI: 10.1007/s10444-004-7614-3.

[3] Xiyuan Hu, Silong Peng and Wen-Liang Hwang. 'EMD Revisited: A New Understanding of the Envelope and Resolving the Mode-Mixing Problem in AM-FM Signals'. In: *IEEE Transactions on Signal Processing* 60.3 (Mar. 2012), pp. 1075–1086. DOI: 10.1109/TSP.2011.2179650.

[4] Boqiang Huang and Angela Kunoth. 'An optimization based empirical mode decomposition scheme'. In: *Journal of Computational and Applied Mathematics* 240 (Mar. 2013), pp. 174–183. DOI: 10.1016/j.cam.2012.07.012.

[5] Xiao-dong Niu, Li-rong Lu, Jian Wang, Xing-cheng Han, Xuan Li and Li-ming Wang. 'An Improved Empirical Mode Decomposition Based on Local Integral Mean and Its Application in Signal Processing'. In: *Mathematical Problems in Engineering* 2021 (Feb. 2021). DOI: 10.1155/2021/8891217.

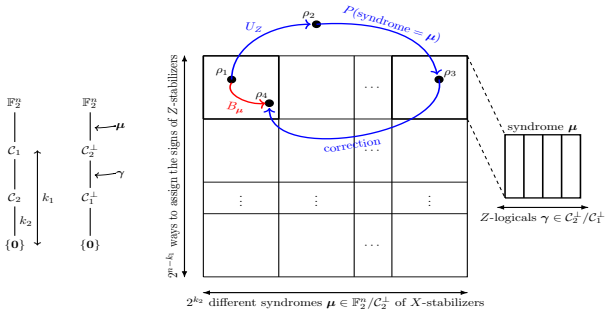
Designing the Quantum Channels Induced by Diagonal Gates

ROBERT CALDERBANK
(joint work with Jingzhen Hu, Qingzhong Liang)

The challenge of quantum computing is to combine error resilience with universal computation. There are many finite sets of gates that are universal, and a standard choice is to augment the set of Clifford gates by a non-Clifford unitary [1]

such as the $T = Z^{1/4}$ gate. Gottesman and Chuang [2] defined the *Clifford hierarchy* when introducing the teleportation model of quantum computing. The first level is the *Pauli group*. The second level is the *Clifford group*, which consists of unitary operators that normalize the Pauli group. The l^{th} level consists of unitary operators that map Pauli operators to the $(l-1)^{\text{th}}$ level under conjugation. The structure of the Clifford hierarchy has been studied extensively [3–8]. For $l \geq 3$, the operators at level l are not closed under matrix multiplication. However, the diagonal gates at each level l of the hierarchy do form a group, and the gates $Z^{1/2^{l-1}}$, $\mathrm{C}^{(i)}Z^{1/2^j}$ with $i + j = l - 1$ generate this group [3, 6].

Quantum error-correcting codes encode logical qubits into physical qubits, and protect information as it is transformed by logical gates. Given a logical diagonal operator among the generators of the diagonal Clifford hierarchy, we describe a general method for synthesizing a CSS code [9,10] preserved by a diagonal physical gate which induces the target logical operator. Logical diagonal gates play a central role in quantum algorithms. In the Shor factoring algorithm [11, 12], our method applies to the $\mathrm{C}^{(i)}Z^{1/2^j}$ diagonal gates which play an essential role in period finding. In magic state distillation (MSD) [13–22], the effectiveness of the protocol depends on engineering the interaction of a diagonal physical gate with the code states of a stabilizer code [23, 24]. Our method transforms a CSS code supporting a lower level logical operator to a CSS code supporting a higher level logical operator. The coefficients in the Pauli expansion of a diagonal gate satisfy a recursion that makes it possible to work backwards from a target logical gate.

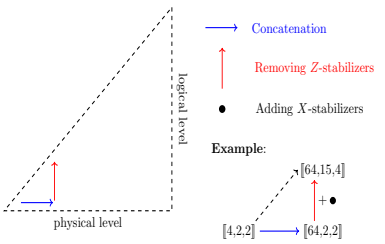


Our approach makes use of an explicit representation of the logical channel induced by a diagonal physical gate. The 2^{n-k_1} rows of the array are indexed by the $[\![n, k_1 - k_2, d]\!]$ CSS codes corresponding to all possible signs of the Z-stabilizer group. The 2^{k_2} columns of the array are indexed by all possible X-syndromes $\boldsymbol{\mu}$. The logical operator $B_{\boldsymbol{\mu}}$ is induced by (1) preparing any code state ρ_1; (2) applying a diagonal physical gate U_Z to obtain ρ_2; (3) using X-stabilizers to measure ρ_2, obtaining the syndrome $\boldsymbol{\mu}$ with probability $p_{\boldsymbol{\mu}}$, and the post-measurement state ρ_3; (4) applying a Pauli correction to ρ_3, obtaining ρ_4. For each syndrome, we expand the induced logical operator in the Pauli basis to obtain the *generator coefficients* [25] that capture state evolution. The generator coefficients $A_{\boldsymbol{\mu},\boldsymbol{\gamma}}$ are obtained by expanding the logical operator $B_{\boldsymbol{\mu}}$ in terms of Z-logical Pauli operators $\epsilon_{(\mathbf{0},\boldsymbol{\gamma})}E(\mathbf{0}, \boldsymbol{\gamma})$, where $\epsilon_{(\mathbf{0},\boldsymbol{\gamma})} \in \{\pm 1\}$. Intuitively, the diagonal physical gate preserves the code space if and only if the induced logical operator corresponding to the trivial syndrome is unitary. To support the objective of fault

tolerance, we emphasize *transversal* gates [23], which are tensor products of unitaries on individual code blocks. The approach taken in prior work is to focus on the code states, and to derive sufficient conditions for a stabilizer code to be fixed by a transversal Z-rotation [13–15, 17–20, 22, 26]. In contrast we derive necessary and sufficient conditions by analyzing the action of a transversal diagonal gate on the stabilizer group that determines the code. An advantage of our approach is that we keep track of the induced logical operator.

The action of a diagonal physical operator U_Z on code states depends very strongly on the signs of Z-stabilizers [25,27,28] and our generator coefficient framework captures how these signs change the logical operators induced by U_Z. For the coherent noise model, a judicious choice of signs creates a decoherence-free subspace, that enables data storage. We demonstrate how to switch between computation and storage by simply applying a Pauli matrix.

Haah [26] used divisibility properties of classical codes to construct CSS codes with parameters $[\![O(d^{l-1}), \Omega(d), d]\!]$ that realize a transversal logical $Z^{1/2^{l-1}}$. Modulo Clifford gates, his construction includes the $[\![2^l, 1, 3]\!]$ punctured quantum Reed-Muller (QRM) codes [18] that support a single logical $Z^{1/2^{l-2}}$ gate, and the family of $[\![6k+8, 2k, 2]\!]$ *triorthogonal* code [15] that support a logical transversal T gate. In contrast we introduce three basic operations - concatenation, removal of Z-stabilizers, and addition of X-stabilizers - that can be combined to synthesize an arbitrary logical diagonal gate [29], and present the $[\![2^m, \binom{m}{r}, 2^{\min\{r,m-r\}}]\!]$ QRM code family [25, 30] as a proof of concept. We also characterize *all* CSS codes with positive signs, invariant under transversal Z-rotation through $\pi/2^l$, that are constructed from classical Reed-Muller (RM) codes by deriving *necessary and sufficient* conditions that relate l to the parameters of the component RM codes [25].



References

[1] P. O. Boykin, T. Mor, M. Pulver, V. Roychowdhury, and F. Vatan, "On universal and fault-tolerant quantum computing," *arXiv preprint quant-ph/9906054*, 1999.

[2] D. Gottesman and I. L. Chuang, "Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations," *Nature*, vol. 402, no. 6760, pp. 390–393, 1999.

[3] B. Zeng, X. Chen, and I. L. Chuang, "Semi-Clifford operations, structure of $\mathcal{C}_k$ hierarchy, and gate complexity for fault-tolerant quantum computation," *Phys. Rev. A*, vol. 77, no. 4, p. 042313, 2008.

[4] S. Beigi and P. W. Shor, "$\mathcal{C}_3$, semi-Clifford and generalized semi-Clifford operations," *arXiv preprint arXiv:0810.5108*, 2008.

[5] I. Bengtsson, K. Blanchfield, E. Campbell, and M. Howard, "Order 3 symmetry in the Clifford hierarchy," *J. Phys. A Math*, vol. 47, no. 45, p. 455302, 2014.

[6] S. X. Cui, D. Gottesman, and A. Krishna, "Diagonal gates in the Clifford hierarchy," *Phys. Rev. A*, vol. 95, no. 1, p. 012329, 2017.

[7] N. Rengaswamy, R. Calderbank, and H. D. Pfister, "Unifying the Clifford hierarchy via symmetric matrices over rings," *Phys. Rev. A*, vol. 100, no. 2, p. 022304, 2019.

[8] T. Pllaha, N. Rengaswamy, O. Tirkkonen, and R. Calderbank, "Un-weyl-ing the Clifford hierarchy," *Quantum*, vol. 4, p. 370, 2020.

[9] A. R. Calderbank and P. W. Shor, "Good quantum error-correcting codes exist," *Phys. Rev. A*, vol. 54, pp. 1098–1105, Aug 1996.

[10] A. M. Steane, "Simple quantum error-correcting codes," *Phys. Rev. A*, vol. 54, no. 6, pp. 4741–4751, 1996.

[11] P. W. Shor, "Algorithms for quantum computation: discrete logarithms and factoring," in *Proc. - Annu. IEEE Symp. Found. Comput. Sci. FOCS*. Ieee, 1994, pp. 124–134.

[12] ——, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM review*, vol. 41, no. 2, pp. 303–332, 1999.

[13] S. Bravyi and A. Kitaev, "Universal quantum computation with ideal Clifford gates and noisy ancillas," *Phys. Rev. A*, vol. 71, no. 2, p. 022316, 2005.

[14] B. W. Reichardt, "Quantum universality from magic states distillation applied to css codes," *Quantum Inf. Process*, vol. 4, no. 3, pp. 251–264, 2005.

[15] S. Bravyi and J. Haah, "Magic-state distillation with low overhead," *Phys. Rev. A*, vol. 86, no. 5, p. 052329, 2012.

[16] H. Anwar, E. T. Campbell, and D. E. Browne, "Qutrit magic state distillation," *New J. Phys.*, vol. 14, no. 6, p. 063006, 2012.

[17] E. T. Campbell, H. Anwar, and D. E. Browne, "Magic-state distillation in all prime dimensions using quantum Reed-Muller codes," *Phys. Rev. X*, vol. 2, no. 4, p. 041021, 2012.

[18] A. J. Landahl and C. Cesare, "Complex instruction set computing architecture for performing accurate quantum z rotations with less magic," *arXiv preprint arXiv:1302.3240*, 2013.

[19] E. T. Campbell and M. Howard, "Unified framework for magic state distillation and multiqubit gate synthesis with reduced resource cost," *Phys. Rev. A*, vol. 95, no. 2, p. 022316, 2017.

[20] J. Haah and M. B. Hastings, "Codes and protocols for distilling t, controlled-s, and toffoli gates," *Quantum*, vol. 2, p. 71, 2018.

[21] A. Krishna and J.-P. Tillich, "Towards low overhead magic state distillation," *Phys. Rev. Lett.*, vol. 123, no. 7, p. 070507, 2019.

[22] C. Vuillot and N. P. Breuckmann, "Quantum pin codes," *arXiv preprint arXiv:1906.11394*, 2019.

[23] D. Gottesman, *Stabilizer codes and quantum error correction*. California Institute of Technology, 1997.

[24] A. R. Calderbank, E. M. Rains, P. Shor, and N. J. Sloane, "Quantum error correction via codes over GF(4)," *IEEE Trans. Inf. Theory*, vol. 44, no. 4, pp. 1369–1387, 1998.

[25] J. Hu, Q. Liang, and R. Calderbank, "Designing the quantum channels induced by diagonal gates," *arXiv preprint arXiv:2109.13481*, 2021.

[26] J. Haah, "Towers of generalized divisible quantum codes," *Phys. Rev. A*, vol. 97, no. 4, p. 042327, 2018.

[27] J. Hu, Q. Liang, N. Rengaswamy, and R. Calderbank, "Mitigating coherent noise by balancing weight-2 Z-stabilizers," *IEEE Trans. Inf. Theory*, pp. 1–1.

[28] D. M. Debroy, L. Egan, C. Noel, A. Risinger, D. Zhu, D. Biswas, M. Cetina, C. Monroe, and K. R. Brown, "Optimizing stabilizer parities for improved logical qubit memories," *arXiv preprint arXiv:2105.05068*, 2021.

[29] J. Hu, Q. Liang, and R. Calderbank, "Climbing the diagonal clifford hierarchy," *arXiv preprint arXiv:2110.11923*, 2021.

[30] N. Rengaswamy, R. Calderbank, M. Newman, and H. D. Pfister, "On optimality of CSS codes for transversal T," *IEEE J. Sel. Areas in Inf. Theory*, vol. 1, no. 2, pp. 499–514, 2020.

Potential and Limitations of Neural Networks for Recovery of Sparse Signals

Felix Krahmer
(joint work with Stefan Bamberger, Reinhard Heckel)

1. Introduction

The field of compressed sensing was introduced in the seminal works [1, 2] and has grown to an intensely studied theory. Its key idea is to enable reconstruction of signals and images from few measurements by imposing constraints on their structure. A commonly studied example of such a structure is sparsity of the signal vectors in some basis. That is, one assumes a signal or image that is well-approximated by a linear combination of just few basis vectors.

In many scenarios, such a signal or image can then be provably recovered from the measurements by solving a convex optimization problem. Compressive sensing enables accelerated magnetic resonance imaging, accelerated computed tomography, and many other applications.

More recently, neural networks have also been empirically demonstrated to exhibit excellent reconstruction performance for such applications and to even outperform classical optimization based methods for a variety of signal and image reconstruction problems. However, contrary to optimization-based methods for which a rich literature on performance guarantees exists [3], for neural network based signal reconstruction, many underlying theoretical questions are still open.

In mathematical terms, the goal of the compressed sensing problem studied in this note is to recover a signal $x \in \mathbb{R}^n$ from m linear measurements $y = Ax \in \mathbb{R}^m$ ($A \in \mathbb{R}^{m\times n}$), for a ground truth signal x that is s-sparse, i.e. at most s of its n entries are non-zero.

1.1. Goal of this work.

Due to the success of neural networks for classification but also certain inverse problems such as the ones arising from MRI scanning, the question arises whether the classical compressed sensing problem of recovering sparse signals can be solved using neural networks. Given measurements $y = Ax$ like above for a sparse signal x, an end-to-end neural network function f should recover x from a coarse approximation given by $A^T y$ such that $x = f(A^T y)$ or at least $\|x - f(A^T y)\|$ is small.

Note that in the first step, a neural network applies a linear transformation to its input. Thus, given f, we can define a network function $\tilde{f}$ such that $\tilde{f}(y) = f(A^T y)$. Considering in addition that $(A^\dagger)^T A^T A = A$ (with $A^\dagger$ being the Moore-Penrose pseudoinverse), we can see that the problems of finding f such that $f(A^T y)$ is (approximately) x and finding $\tilde{f}$ such that $\tilde{f}(y)$ is (approximately) x are equivalent. This is why we consider the latter case in the results of this work.

The goal of this work is to investigate for what kinds of neural networks it is possible to reconstruct any sparse signal x from its measurements. For a negative result, we restrict this analysis to the case of 1-sparse vectors. Since impossibility of recovering 1-sparse vectors also implies impossibility of recovering vectors of

any higher sparsity level, this is not an essential restriction in this direction. In all results, we assume that all networks have the commonly used ReLU activation function given by $\mathrm{ReLU}(x) = \max(0, x)$ and use specific properties of it.

Specifically, we show that no ReLU -neural network with only one hidden layer can recover all 1-sparse vectors in the setup of interest, regardless of the size of the network. We also show that there exist relu-networks with two hidden layers that can perform approximate recovery with an arbitrary precision. Moreover, we show that the previously developed tools can be applied to show that exact reconstruction of s-sparse vectors is possible with $\lceil \log_2 s + 2 \rceil$ layers. However, in this case of exact recovery, it is unknown whether the shown number of layers is optimal.

2. Main Results

The main results of this work are the following two theorems. Theorem 2.1 states that a ReLU network with one hidden layer cannot recover all s-sparse vectors from any $m \ll n$ linear measurements, not even approximately and for 1-sparse vectors. In strong contrast, Theorem 2.2 states that for a ReLU network with two hidden layers, recovery of all s-sparse vectors is possible to an arbitrary precision $\epsilon > 0$ under weak assumptions on the matrix A.

Theorem 2.1 (Impossibility result for one hidden layer)**.** *Let $A \in \mathbb{R}^{m\times n}$, $m \leq n$, and $f : \mathbb{R}^m \to \mathbb{R}^n$ be a function described by a (possibly biased) neural network with one hidden layer,* ReLU *activation function, and arbitrary width. Then,*

$$\sup_{x\in\Sigma_1\setminus\{0\}} \frac{\|x - f(Ax)\|_2}{\|x\|_2} \geq \sqrt{1 - \frac{m}{n}}.$$

Note that in the case of underdetermined recovery problems, arguably the case of particular interest, $m \ll n$ such that the lower bound for the relative error is close to 1.

Theorem 2.2 (Arbitrary precision recovery with two hidden layers)**.** *Let $s \geq 1$ and $A \in \mathbb{R}^{m\times n}$ be a matrix for which $Ax \neq 0$ for all $2s$-sparse $x \in \mathbb{R}^n$. Let $\epsilon \in (0, 1)$. There exists a function $f : \mathbb{R}^m \to \mathbb{R}^n$, represented by an unbiased* ReLU *network with two hidden layers such that for all s-sparse $x \in \mathbb{R}^n$,*

$$\sup_{x\in\Sigma_s\setminus\{0\}} \frac{\|x - f(Ax)\|_2}{\|x\|_2} \leq \epsilon.$$

With the same methods as Theorem 2.2, also the following result about exact recovery can be shown.

Theorem 2.3 (Exact recovery for sparsity s)**.** *Let $A \in \mathbb{R}^{m\times n}$ be a matrix for which $Ax \neq 0$ for all $2s$-sparse $x \in \mathbb{R}^n$, $s \geq 2$. There exists a function $f : \mathbb{R}^m \to \mathbb{R}^n$, represented by an unbiased* ReLU *network with $\lceil \log_2(s) \rceil + 2$ hidden layers such that $f(Ax) = x$ for all s-sparse $x \in \mathbb{R}^n$.*

A key ingredient to this theorem is a general result about the exact representation of continuous piecewise linear functions by ReLU networks [4].

Theorems 2.1 and 2.2 show that for solving the sparse recovery problem to an arbitrary precision with ReLU networks, two hidden layers are necessary and sufficient. In contrast, the universal approximation theorem [5] states that one hidden layer is enough to approximate any continuous function on a compact domain. So the key difference that the domain of the sparse recovery problem, that is the set of all images of s-sparse vectors under A, is not compact.

Previous approaches of unfolding the iterative shrinkage thresholding algorithm (ISTA) had already shown the possibility of approximately solving the sparse recovery problem with sparsity s for networks of depth $\mathcal{O}(\log s)$ [6]. Theorem 2.2 improves this to precisely two hidden layers and Theorem 2.3 shows that the former depth is even sufficient for exact recovery.

References

[1] E J. Candès, J. Romberg, and T Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., **59** (2006), 1207–1223.

[2] D. L. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory, **52** (2006), 1289–1306.

[3] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[4] Juncai He et al., ReLU *deep neural networks and linear finite elements*, Journal of Computational Mathematics 38.3 (2020), pp. 502–527. issn: 0254-9409.

[5] G. Cybenko. *Approximation by Superpositions of a Sigmoidal Function*, Mathematics of Control, Signals, and Systems 2 (1989), pp. 303–314.

[6] Xiaohan Chen et al. *Theoretical Linear Convergence of Unfolded ISTA and Its Practical Weights and Thresholds*, Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS'18. Montreal, Canada: Curran Associates Inc., 2018, 9079–9089.

The role of recurrence and stochasticity in learning streaming data

Youness Boutaib

(joint work with Wiebke Bartolomaeus, Sandra Nestler, Holger Rauhut)

Recurrent neural networks (RNNs) constitute the simplest machine learning paradigm that is able to handle variable-length data sequences while tracking long term dependencies and taking into account the temporal order of the received information. These data streams appear naturally in many fields such as signal processing or financial data. The RNN architecture is inspired from biological neural networks where both recurrent connectivity and stochasticity in the temporal dynamics are ubiquitous. Despite the empirical success of RNNs and their many variants (long short-term memory networks (LSTMs), gated recurrent units (GRUs), etc.), several fundamental mathematical questions related to the functioning of these networks remain open:

- What is the exact type of information that an RNN learns from the input sequences?
- Training artificial RNNs with classical methods like gradient descent suffer from fundamental problems such as instability, non-convergence, exploding gradient errors [2] and plateauing [4]. On the other hand, biological networks seem to be robust and easy to train. How does stochasticity contribute in regard to this?
- What is the amount of data needed for such a network to achieve a small estimation error with high probability?

We set out to answer these questions by modelling a biological neural network as a continuous-time (stochastic) RNN with a randomly chosen connectivity matrix and an identity activation function in view of classifying data streams:

- The continuous-time dynamics provide us with a richer mathematical toolbox while keeping key features and issues of such systems such as the dependence on the whole data sequence and its order.
- Randomly generating the connectivity matrix of an RNN is the cornerstone of reservoir computing [3, 5] which has shown exceptional performances in a variety of tasks.
- We choose to work with identity activation functions in order to build the intuition as to the answer to the questions above. In this case, we obtain precise formulas. We aim to generalise the results of this study to the non-linear case in a later work.

More explicitly, the input and the hidden state of the RNN are modelled, respectively, as high-dimensional time-dependent continuous paths x and y. The dynamics of the latter are dictated by the following stochastic differential equation (SDE):

$$\mathrm{d}y_t = (-y_t + Wy_t + u(x_t))\mathrm{d}t + \Sigma \mathrm{d}B_t, \quad t \leq T. \tag{1}$$

Here, u is a pre-processing map, W is the network matrix that models the connection strength between neurons and Σ describes the random effect of a Brownian noise B. In line with most common practices, a hyperplane classifier h is combined with the final hidden state of the neural network to produce a prediction $v = h(y_T(x))$.

In [1], we approach the binary classification problem from the point of view of statistical learning theory. Note that as the hidden state vector $y_T(x)$ is random, the global hypothesis class $\mathcal{H}$ is itself a class of random learners, thus prompting an adjustment of the learning setup and proof techniques. We give a generalisation error bound that shows that minimising the empirical risk achieves agnostic PAC learnability and gives guarantees on the ability of the empirical risk minimiser to generalise to unseen data. Consequently, we study in details the empirical risk minimisation (ERM) procedure:

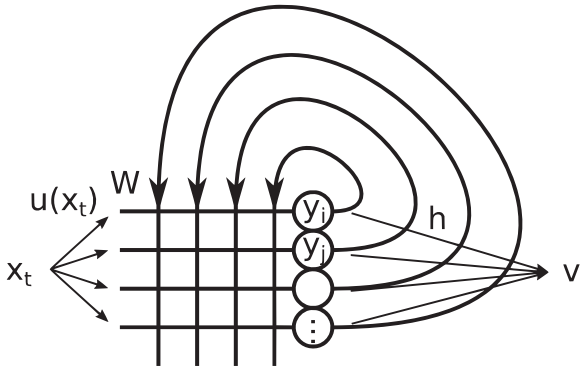


FIGURE 1. A sketch of the recurrent learning architecture

- we compare its output to that of the popular Support Vector Machine (SVM). In particular, the solution to the ERM algorithm is sensitive to the number of inputs in each class (thus, in some way, "learning" the data generating distribution), but stable against outliers and does not suffer from non-convergence problems encountered by SVM algorithms when combined with gradient descent to produce an optimal pre-processing map u.
- We argue heuristically that noise, which is a natural assumption in modelling biological neural networks, provides stability and robustness against different types of perturbations to the dataset.
- We show rigorously that in the linear case, the RNN retains a "partial signature" of the time-lifted input signal as global information about said signal. The empirical risk is a function of the tunable parameters of the model and the partial signatures of the training data $S(x) = \left(\int_0^T \frac{(T-s)^k}{k!} x_s \mathrm{d}s\right)_{k\geq 0}$. This means that the RNN can learn the classification task based only on the partial signature S of the training data and that it cannot distinguish between two paths with the same partial signatures.

Finally, we look into the numerical minimisation of the empirical risk using gradient descent. The experiments are performed using the Japanese vowel dataset[1] (12-dimensional paths) and classes of 5-dimensional trigonometric polynomials.

REFERENCES

[1] Wiebke Bartolomaeus, Youness Boutaib, Sandra Nestler, and Holger Rauhut. Path classification by stochastic linear recurrent neural networks, 2021.

[2] Kenji Doya. Bifurcations in the learning of recurrent neural networks 3. *learning (RTRL)*, 3:17, 1992.

[1] https://archive.ics.uci.edu/ml/datasets/Japanese+Vowels

[3] Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *science*, 304(5667):78–80, 2004.
[4] Zhong Li, Jiequn Han, Qianxiao Li, et al. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. *arXiv preprint arXiv:2009.07799*, 2020.
[5] Wolfgang Maass, Prashant Joshi, and Eduardo D Sontag. Computational aspects of feedback in neural circuits. *PLoS Comput Biol*, 3(1):e165, 2007.

Generalization in Deep Learning Through the Lens of Implicit Rank Minimization

Noam Razin
(joint work with Asaf Maman, Nadav Cohen)

One of the central mysteries in deep learning is the ability of overparameterized neural networks to generalize, even in the absence of any explicit regularization. Conventional wisdom is that gradient-based optimization induces an *implicit regularization* — a tendency to fit training examples with predictors of minimal "complexity." Mathematically characterizing this tendency is a major open problem in the theory of deep learning.

A widespread hope is that a characterization based on minimization of norms may apply, and a standard testbed for studying this prospect is matrix factorization — matrix completion via linear neural networks. It was an open question whether norms can explain the implicit regularization in matrix factorization. Though initially it was conjectured that the nuclear norm is implicitly minimized ([1]), in [2] we resolve this open question in the negative, by proving that there exist natural matrix factorization problems on which the implicit regularization drives *all* norms *towards infinity* while minimizing rank. This suggests that, rather than perceiving the implicit regularization via norms, a potentially more useful interpretation is minimization of rank.

The natural question that arises is whether the interpretation of implicit rank minimization applies to more practical settings. In [3] and [4] we show that the tendency to low rank extends from matrices to tensors (multi-dimensional arrays). Namely, by characterizing the dynamics that gradient descent induces on tensor factorizations, we establish that these result in incremental learning, creating a bias towards low *tensor ranks*. Analogously to how matrix factorization can be viewed as a linear neural network, tensor factorizations correspond to a class of non-linear convolutional networks, for which low tensor ranks implies a bias towards local dependencies. Motivated by this observation, we empirically demonstrate that: *(i)* simple natural image datasets (MNIST and FMNIST) are fittable with predictors of extremely low tensor ranks, explaining why generalization on such datasets is achieved; *(ii)* other image datasets, which entail strong dependence between spatially distant pixels, lead convolutional networks to completely fail in generalizing; and *(iii)* it is possible to greatly improve generalization by employing dedicated explicit regularization which promotes high tensor ranks, *i.e.* long range dependencies.

Overall, our results suggest that notions of rank (such as tensor ranks) may shed light on both implicit regularization of neural networks, and properties of real-world data translating this implicit regularization to generalization.

References

[1] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Implicit regularization in matrix factorization*, Advances in Neural Information Processing Systems, 2017.

[2] N. Razin and N. Cohen, *Implicit regularization in deep learning may not be explainable by norms*, Advances in Neural Information Processing Systems, 2020.

[3] N. Razin, A. Maman, and N. Cohen, *Implicit regularization in tensor factorization*, International Conference on Machine Learning, 2021.

[4] N. Razin, A. Maman, and N. Cohen, *Analyzing implicit regularization in convolutional networks via hierarchical tensor factorization*, work in progress (WIP), 2021.

Riesz bases of exponentials for partitions of intervals

Götz Pfander

(joint work with Shauna Revay, David Walnut)

Fourier series form a cornerstone of analysis; it allows the expansion of $L^2[0,1]$ functions in the orthonormal basis of integer frequency exponentials $\mathcal{E}(\mathbb{Z}) = \{e^{2\pi ikx}\}_{k\in\mathbb{Z}}$. A simple rescaling argument shows that by splitting the integers into evens and odds, we obtain orthogonal bases for functions defined on the first, respectively the second half of the unit interval, that is, $\mathcal{E}(2\mathbb{Z})$ is an orthogonal bases of $L^2[0,1/2]$ and $\mathcal{E}(2\mathbb{Z}+1)$ is an orthogonal bases of $L^2[1/2,1]$.[1]

Building on this curiosity, we explain that, given any finite partition of the unit interval into subintervals, we can split the integers into subsets, each of which forms a Riesz basis (not necessarily orthogonal) for functions on the respective subinterval [5–7]. (The case of 2 subintervals was considered by Kristian Seip in [8].)

Theorem 1 ([5]). *For $a_0 = 0 < a_1 < a_2 < \ldots < a_n = 1$ exist pairwise disjoint sets $\Lambda_1, \Lambda_2, \ldots, \Lambda_n \subseteq \mathbb{Z}$ with $\Lambda_1 \cup \Lambda_2 \cup \ldots \cup \Lambda_n = \mathbb{Z}$ so that $\mathcal{E}(\Lambda_k)$ is a Riesz basis for $L^2[a_{k-1}, a_k]$.*

This result was then generalized to provide hierarchical Riesz bases with integer frequencies as follows.

Theorem 2 ([6]). *For $b_1, \ldots, b_n > 0$ with $\sum_{j=1}^n b_j = 1$ exist pairwise disjoint sets $\Lambda_1, \ldots, \Lambda_n \subseteq \mathbb{Z}$ with $\bigcup_{j=1}^n \Lambda_j = \mathbb{Z}$ and the property that $\mathcal{E}\big(\bigcup_{j\in J} \Lambda_i\big)$ is a Riesz basis for $L^2(I)$ for I any interval of length $\sum_{j\in J} b_j$ for any $J \subseteq \{1, \ldots, n\}$.*

In the countable setting, we obtain

[1] Here and in the following, we shall use the customary notation $\mathcal{E}(\Lambda) = \{e^{2\pi i\lambda x}\}_{\lambda\in\Lambda}$.

Theorem 3 ([6]). *Let $b_1, b_2, \ldots > 0$ satisfy $\sum_{j=1}^{\infty} b_j = 1$. For fixed $K \in \mathbb{N}$ there exist pairwise disjoint sets $\Lambda_1, \Lambda_2, \ldots \subseteq \mathbb{Z}$ with the property that for any $J \subseteq \mathcal{N}$ with $|J| \leq K$ we have $\mathcal{E}\big(\bigcup_{j\in J} \Lambda_j\big)$ is a Riesz basis for $L^2(I)$ for I any interval of length $\sum_{j\in J} b_j$.*

It is worth to emphasize that Riesz bases of exponentials can not be combined as above in general, that is, if $\mathcal{E}(\Lambda_1)$ and $\mathcal{E}(\Lambda_2)$ form Riesz bases for $L^2(I_1)$ and $L^2(I_2)$ respectively, it need not follow that $\mathcal{E}(\Lambda_1 \cup \Lambda_2)$ forms a Riesz basis for $L^2(I_1 \cup I_2)$, even if Λ_1 and Λ_2 are disjoint.

Our results are based on combining tools from analysis, probability and number theory. Key is the development of an Avdonin map calculus that allows for an iterative use of Avdonin's theorem [1]:

Theorem 4. *For $\varphi : \frac{\mathbb{Z}+\alpha}{a} \to \mathbb{R}$ injective with separated range, $\mathcal{E}\big(\mathrm{Range}\,\varphi\big)$ is a Riesz basis for $L^2[0,a]$ if for some $R > 0$ it holds*

$$\sup_{m\in\mathbb{Z}} \left| \frac{1}{R} \sum_{k\in[mR,(m+1)R)} \varphi\big(\tfrac{k+\alpha}{a}\big) - \tfrac{k+\alpha}{a} \right| < \frac{1}{4a}.$$

This is combined with a suited formulation of Weyl-Khinchin equidistribution theorem

Theorem 5. *For a irrational and $\epsilon > 0$ exists N_0 such that for $N \geq N_0$ and $m \in \mathbb{Z}$,*

$$\left| \frac{1}{N} \sum_{k=mN}^{(m+1)N-1} \frac{k+\frac{1}{2}}{a} \mod 1 \quad - \frac{1}{2} \right| < \epsilon.$$

as well as with a result on inhomogeneous Beatty sequences [2–4]

Theorem 6. *For a, b irrational with $a + b = 1$, the sets $\left\{ \left[\frac{k+\frac{1}{2}}{a} \right]_{\mathbb{Z}+\frac{1}{2}} \right\}_{k\in\mathbb{Z}}$ and $\left\{ \left[\frac{\ell+\frac{1}{2}}{b} \right]_{\mathbb{Z}+\frac{1}{2}} \right\}_{\ell\in\mathbb{Z}}$ partition $\mathbb{Z}+\frac{1}{2}$, where $[x]_K$ rounds x to the nearest element in K, choosing the lesser if two elements in K have the same distance to x.*

For details we refer to [6].

References

[1] S. Avdonin. On the question of Riesz bases of complex exponential function in l^2. *Vestnik Leningrad Univ. Ser. Mat.*, 13:5–12, 1974.

[2] Samuel Beatty. Problems and Solutions: Problems for Solutions: 3173. *Amer. Math. Monthly*, 33(3):159, 1926.

[3] Samuel Beatty, A. Ostrowski, J. Hyslop, and A. C. Aitken. Problems and Solutions: Solutions: 3177. *Amer. Math. Monthly*, 34(3):159–160, 1927.

[4] Aviezri S. Fraenkel. The bracket function and complementary sets of integers. *Canadian J. Math.*, 21:6–27, 1969.

[5] Götz E. Pfander, Shauna Revay, and David Walnut. Riesz bases of exponentials for partitions of intervals. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–4, 2019.

[6] Götz E. Pfander, Shauna Revay, and David Walnut. Exponential bases for partitions of intervals. preprint, arXiv:2109.04441.

[7] Shauna Revay. *Unions of Riesz bases of exponentials for bandlimited signals*. PhD thesis, George Mason University, Department of Mathematical Sciences, Fairfax, VA 22030, May 2018.

[8] Kristian Seip. A simple construction of exponential bases in L^2 of the union of several intervals. *Proc. Edinburgh Math. Soc. (2)*, 38(1):171–177, 1995.

Reporter: Youness Boutaib

Participants

Prof. Dr. Rima Alaifari
Departement Mathematik
ETH-Zentrum
HG G 59.2
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Bubacarr Bah
African Institute for Mathematical Sciences - Center of Excellence in Mathematical Finance
6-8 Melrose Road
Muizenberg 7945
SOUTH AFRICA

Prof. Dr. Afonso S. Bandeira
Departement Mathematik
ETH-Zentrum
Rämistr. 101
8092 Zürich
SWITZERLAND

Prof. Dr. Holger Boche
LST für Theoretische Informationstechnik
Technische Universität München (LTI)
Theresienstrasse 90/IV
80333 München
GERMANY

Prof. Dr. Bernhard G. Bodmann
Department of Mathematics
University of Houston
Houston TX 77204-3008
UNITED STATES

Dr. March Boedihardjo
Department of Mathematics
University of California, Irvine
Irvine, CA 92697-3875
UNITED STATES

Prof. Dr. Helmut Bölcskei
Mathematical Information Sciences
ETH Zürich
Room: ETF E 122
Sternwartstrasse 7
8092 Zürich
SWITZERLAND

Dr. Youness Boutaib
Mathematics for Information Processing
RWTH Aachen
Pontdriesch 10
52062 Aachen
GERMANY

Prof. Dr. Joan Bruna
Courant Institute of Mathematical Sciences
New York University
251, Mercer Street
New York, NY 10012-1110
UNITED STATES

Prof. Dr. A. Robert Calderbank
Gross Hall, Room 317
Pratt School of Engineering
Duke University
140 Science Drive
P.O. Box 90984
27708 Durham, NC 27708
UNITED STATES

Prof. Dr. Xiuyuan Cheng
Department of Mathematics
Duke University
P.O.Box 90320
Durham, NC 27708-0320
UNITED STATES

Prof. Dr. Alexander Cloninger
Department of Mathematics
Yale University
Box 208 283
New Haven, CT 06520
UNITED STATES

Prof. Dr. Albert Cohen
Laboratoire Jacques-Louis Lions
Sorbonne Université
4, Place Jussieu
75005 Paris Cedex
FRANCE

Dr. Nadav Cohen
Department of Mathematics
School of Mathematical Sciences
Tel Aviv University
Ramat Aviv, Tel Aviv 69978
ISRAEL

Prof. Dr. Wolfgang Dahmen
Department of Mathematics
University of South Carolina
1523 Greene Street
Columbia, SC 29208
UNITED STATES

Prof. Dr. Ingrid Daubechies
Department of Mathematics
Duke University
P.O.Box 90320
Durham, NC 27708-0320
UNITED STATES

Prof. Dr. Christine De Mol
Department of Mathematics
Université Libre de Bruxelles
CP 217 Campus Plaine
Boulevard du Triomphe
1050 Bruxelles
BELGIUM

Dr. Sjoerd Dirksen
Mathematisch Instituut
Universiteit Utrecht
POB 80.010
Budapestlaan 6
3508 TA Utrecht
NETHERLANDS

Dr. Markus Faulhuber
Fakultät für Mathematik
Universität Wien
Oskar-Morgenstern-Platz 1
1090 Wien
AUSTRIA

Prof. Dr. Hans Georg Feichtinger
Fakultät für Mathematik
Universität Wien
Oskar-Morgenstern-Platz 1
1090 Wien
AUSTRIA

Dr. Leonardo Galli
Mathematics for Information Processing
RWTH Aachen
Pontdriesch 10
52062 Aachen
GERMANY

Dr. Martin Genzel
Mathematisch Instituut
Universiteit Utrecht
Budapestlaan 6
P.O. Box P. O. Box 80.01
3508 TA Utrecht
NETHERLANDS

Prof. Dr. Remi Gribonval
Laboratoire de l'Informatique du
Parallélisme
ENS de Lyon et INRIA
46 Allée d'Italie
F-69007 Lyon Cedex
FRANCE

Prof. Dr. Philipp Grohs
Fakultät für Mathematik
Universität Wien
Oskar Morgenstern Platz 1
1090 Wien
AUSTRIA

Prof. Dr. David Groß
Institut für Theoretische Physik
Universität Köln
Zülpicher Straße 77
50937 Köln
GERMANY

Prof. Dr. Matthew Hirn
Michigan State University
Department of Computational
Mathematics, Science & Engineering
Room 2507F, Engineering Building
428 S. Shaw Ln.
East Lansing MI, 48824
UNITED STATES

Frederik Hoppe
Lehrstuhl für Mathematik der
Informationsverarbeitung
RWTH Aachen University
Pontdriesch 10
52062 Aachen
GERMANY

Laslo Hunhold
Department Mathematik/Informatik
Abteilung Mathematik
Universität zu Köln
Weyertal 86-90
50931 Köln
GERMANY

Dr. Shahar Kovalsky
Department of Mathematics
University of North Carolina
at Chapel Hill
Phillips Hall
Chapel Hill NC 27599-3250
UNITED STATES

Prof. Dr. Felix Krahmer
Zentrum Mathematik
Lehr- u. Forschungseinheit M 15
Technische Universität München
Boltzmannstrasse 3
85748 Garching bei München
GERMANY

Prof. Dr. Richard Küng
Department of Computer Science
Johannes Kepler University Linz
SCP3 0405
Altenberger Strasse 69
4040 Linz
AUSTRIA

Prof. Dr. Gitta Kutyniok
Mathematisches Institut
Ludwig-Maximilians-Universität
München
Theresienstraße 39
80333 München
GERMANY

Prof. Dr. Mauro Maggioni
Department of Mathematics
Johns Hopkins University
Baltimore, MD 21218-2689
UNITED STATES

Prof. Dr. Shahar Mendelson
Centre for Mathematics & its
Application
Australian National University
Canberra ACT 0200
AUSTRALIA

Prof. Dr. Hrushikesh N. Mhaskar
Institute of Mathematical Sciences
Claremont Graduate University
Claremont, CA 91711
UNITED STATES

Dr. Dustin G. Mixon
Department of Mathematics
The Ohio State University
100 Mathematics Building
231 West 18th Avenue
Columbus, OH 43210-1174
UNITED STATES

Prof. Dr. Guido F. Montufar
University of California, Los Angeles,
and Max-Planck-Institut für
Mathematik in den Naturwissenschaften
Inselstrasse 22
04103 Leipzig
GERMANY

Prof. Dr. Deanna Needell
Department of Mathematics
University of California at
Los Angeles
Los Angeles, CA 90095-1555
UNITED STATES

Gabin Maxime Nguegnang
Institut für Mathematik
RWTH Aachen
Templergraben 55
52062 Aachen
GERMANY

Laura Paul
Lehrstuhl für Mathematik der
Informationsverarbeitung
RWTH Aachen University
Pontdriesch 10
52062 Aachen
GERMANY

Prof. Dr. Götz Pfander
Mathematisch-Geographische Fakultät
Katholische Universität
Eichstätt-Ingolstadt
Ostenstrasse 26-28
85072 Eichstätt
GERMANY

Prof. Dr. Gerlind Plonka-Hoch
Institut f. Numerische & Angew.
Mathematik
Universität Göttingen
Lotzestrasse 16-18
37083 Göttingen
GERMANY

Prof. Dr. Holger Rauhut
LST Mathematik der
Informationsverarbeitung
RWTH Aachen
Pontdriesch 10
52062 Aachen
GERMANY

Dr. Noam Razin
Department of Mathematics
School of Mathematical Sciences
Tel Aviv University
P.O.Box 39040
Ramat Aviv, Tel Aviv 69978
ISRAEL

Prof. Dr. Karin Schnass
Institut für Mathematik
Universität Innsbruck
Technikerstrasse 13
6020 Innsbruck
AUSTRIA

Ekkehard Schnoor
Institut für Mathematik
RWTH Aachen
Templergraben 55
52062 Aachen
GERMANY

Shan Shan
Mathematics Department
Mount Holyoke College
South Hadley, MA 01075
UNITED STATES

Prof. Dr. Barak Sober
Department of Mathematics
Duke University
P.O.Box 90320
Durham, NC 27708-0320
UNITED STATES

Dr. Mahdi Soltanolkotabi
Ming Hsieh Department of Electrical Engineering
University of Southern California
3740 McClintock Avenue
Los Angeles CA 90089-2565
UNITED STATES

Dr. Stefan Steinerberger
Department of Mathematics
University of Washington, Seattle
Seattle, WA 98195
UNITED STATES

Prof. Dr. Dominik Stöger
Mathematisch-Geographische Fakultät
Katholische Universität Eichstätt
Ostenstraße 26-28
85072 Eichstätt
GERMANY

Prof. Dr. Thomas Strohmer
Department of Mathematics
University of California, Davis
1, Shields Avenue
Davis, CA 95616-8633
UNITED STATES

Laura Thesing
Mathematisches Institut
Ludwig-Maximilians-Universität München
Theresienstraße 39
80333 München
GERMANY

Dr. Soledad Villar
Department of Mathematics
Johns Hopkins University
Baltimore, MD 21218-2689
UNITED STATES

Dr. Felix Voigtlaender
Zentrum Mathematik
Technische Universität München
Boltzmannstrasse 3
85748 Garching bei München
GERMANY

Prof. Dr. Rachel Ward
POB 10.144
Department of Mathematics
University of Texas at Austin
2515 Speedway
Austin, TX 78712
UNITED STATES

Prof. Dr. Hau-Tieng Wu
Department of Mathematics
Duke University
P.O.Box 90320
Durham, NC 27708-0320
UNITED STATES