Mathematisches Forschungsinstitut Oberwolfach

Report No. 14/2018

DOI: 10.4171/OWR/2018/14

Applied Harmonic Analysis and Data Processing

Organised by
Ingrid Daubechies, Durham
Gitta Kutyniok, Berlin
Holger Rauhut, Aachen
Thomas Strohmer, Davis

25 March – 31 March 2018

Abstract. Massive data sets have their own architecture. Each data source has an inherent structure, which we should attempt to detect in order to utilize it for applications, such as denoising, clustering, anomaly detection, knowledge extraction, or classification. Harmonic analysis revolves around creating new structures for decomposition, rearrangement and reconstruction of operators and functions—in other words inventing and exploring new architectures for information and inference. Two previous very successful workshops on applied harmonic analysis and sparse approximation have taken place in 2012 and in 2015. This workshop was the an evolution and continuation of these workshops and intended to bring together world leading experts in applied harmonic analysis, data analysis, optimization, statistics, and machine learning to report on recent developments, and to foster new developments and collaborations.

Mathematics Subject Classification (2010): 42-XX, 65Txx, 94Axx, 65K05, 15A52.

Introduction by the Organisers

The workshop Applied Harmonic Analysis and Data Processing was organized by Ingrid Daubechies (Durham), Gitta Kutyniok (Berlin), Holger Rauhut (Aachen) and Thomas Strohmer (Davis). This meeting was attended by 49 participants from three continents. Advances in technology and the ever-growing role of digital sensors and computers in science have led to an exponential growth in the amount and complexity of data we collect. Uncertainty, scale, non-stationarity, noise, and heterogeneity are fundamental issues impeding progress at all phases of the pipeline that creates knowledge from data. 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 a range of techniques towards the problem of efficiently representing, analyzing, compressing, and processing with "Big Data". Massive data sets have their own architecture. Each data source has an inherent structure, which we should attempt to detect in order to utilize it for applications, such as denoising, clustering, anomaly detection, knowledge extraction, recovery, etc. Harmonic analysis revolves around creating new structures for decomposition, rearrangement and reconstruction of operators and functions—in other words inventing and exploring new architectures for information and inference. Indeed, in the last three decades Applied Harmonic Analysis has been at the center of many significant new ideas and methods crucial in a wide range of signal and image processing applications, and in the analysis and processing of large data sets. For example, compressive sensing, sparse approximations and models, geometric multiscale analysis and diffusion geometry represent some quite recent important breakthroughs.

Several new directions have emerged on the heels of compressive sensing: Low-rank matrix recovery aims at recovering a matrix with small rank from incomplete data. In particular, matrix completion recovers the matrix from only a small fraction of its entries. Since low-rank structures arise in numerous applications, one can expect an enormous impact. However, much of the theory so far deals with linear measurements, while in practice we often also face non-linear measurements, for instance in situations where only signal intensity can be obtained. Despite recent breakthroughs in the area of phase retrieval, many challenging mathematical problems problems remain open in these areas.

Inverse problems arising in connection with massive, complex data sets pose tremendous challenges and require new mathematical tools. Numerous deep questions arise. How can we utilize ideas of sparsity and minimal information complexity in this context? Is there a unified view of such measures that would include sparsity, lowrankness, and others (such as low-entropy), as special cases? This may lead to a new theory that considers an abstract notion of simplicity in general inverse problems. An important emerging topic in this context is the design efficient non-convex algorithms with provable convergence guarantees.

One of the most exciting developments in machine learning in the past five years is the advent of deep learning, which is a special form of a neural network. Deep neural networks, and in particular convolutional networks have recently achieved state-of-the-art results on several complex computer vision and speech recognition tasks. However, until now deep learning acts very much like a black box, since algorithms are often based on ad hoc rules without theoretical foundation, the learned representations lack intepretability; we do not really understand why certain deep networks succeed and and we do not know how to modify them for those cases where they fail. Thus, developing a mathematical foundation for deep

learning is an important and rather challening task in data science, and one part of this workshop was dedicated to this topic.

This workshop was a concerted effort to bring together researchers with various backgrounds, including harmonic analysis, optimization, probabibility theory, group theory, approximation theory, computer science, machine learning, and electrical engineering. The workshop featured 27 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 resarch. At the same time it is very entertaining for the audience. Almost all of the attendees participated, ranging from PhD students to renowned professors, contributing to the success of this session.

Some highlights of the program included:

- **Advanced sampling theory:** One of the problems that link harmonic analysis with data processing is the sampling problem. The main theoretical issue is how the stability of sampling and recovery is related to the number or density of samples. Related issues are the questions of localization, non-uniform sampling, and last not least suitable numerical algorithms. Karlheinz Gröchenig presented a range of compelling results using tools from shift-invariant spaces and totally positive functions. Albert Cohen discussed function approximation from sampling in high dimensions using optimal weighted least squares approximation. Felix Krahmer talked about "unlimited sampling", a mathematical framework for sampling that can overcome limitations in current analog-to-digital converters.
- **Nonlinear inverse problems:** In many applications we can only acquire nonlinear measurements of the function of interest. Phase retrieval is but the most prominent example. Several talks were dedicated to nonlinear inverse problems. Babak Hassibi and Rima Alaifari both presented recent progress in the solution of the phase retrieval problem, while Yuxin Chen and Justin Romberg highlighted exciting progress in convex and nonconvex optimization for certain nonlinear problems.
- **Emerging theory of Deep Learning:** Despite the huge practical successes of Deep Learning in recent years, the mathematical understanding of deep learning is in its infancy. Several talks aimed at to remedy this situation. Philipp Grohs demonstrated how to avoid the curse of dimensionality when solving Kolmogorov equation in high dimensions by means of deep learning. Mahdi Soltanolkoltabi and Remi Gribonval were among several speakers who presented theoretical progress towards understanding some of the heuristics behind neural networks.

The organizers would like to take the opportunity to thank MFO for providing support and a very inspiring environment for the workshop. The magic of the place and the pleasant atmosphere contributed greatly to the success of 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: Applied Harmonic Analysis and Data Processing

Table of Contents

Abstracts

Sampling in Shift-Invariant Spaces, Gabor Frames, and Totally Positive Functions

Karlheinz Gröchenig
(joint work with José Luis Romero, Joachim Stöckler)

One of the problems that link harmonic analysis with data processing is the sampling problem: given data $(x_j, y_j)_{j\in J} \subseteq \mathbb{R}^d \times \mathbb{C}$, one should find or recover or approximate or learn a function f such that $f(x_j) \approx y_j$. In this talk we considered the sampling problem with respect to a given a priori signal model in dimension $d = 1$. We assume that f is contained in a shift-invariant space. Precisely, f lies in a subspace of $L^2(\mathbb{R})$ of the form

$$V(g) = \{f = \sum_{k\in\mathbb{Z}} c_k g(\cdot - k) : c \in \ell^2(\mathbb{Z})\}\,,$$

where g is a fixed, well localized and smooth generating function.

The main theoretical issue is how the stability of sampling and recovery is related to the number or density of samples. Related issues are the questions of localization, non-uniform sampling, and last not least suitable numerical algorithms.

For stability, we say that $\Lambda \subseteq \mathbb{R}$ is a *set of stable sampling* for $V(g) \subseteq L^2(\mathbb{R})$, if there exist $A, B > 0$ (the sampling constants), such that

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda} |f(\lambda)|^2 \leq B\|f\|_2^2 \qquad \forall f \in V(g)\,.$$

To explain how many samples are sufficient to recover a function in $V(g)$, we use the *Beurling density* of $\Lambda \subseteq \mathbb{R}$ defined by

$$D^-(\Lambda) = \liminf_{r\to\infty} \inf_{x\in\mathbb{R}} \frac{\#\Lambda \cap [x, x + r]}{r}$$

The paradigm of sampling theory is the classical theory of bandlimited functions covered by the theorems of Beurling and Landau. In this case the generator is $g(t) = \frac{\sin \pi x}{\pi x}$ or $\hat{g} = \chi_{[-1/2,1/2]}$ and the corresponding shift-invariant space $V(g)$ coincides with the Paley-Wiener space

$$\mathcal{B} = \{f \in L^2(\mathbb{R}) : \operatorname{supp} \hat{f} \subseteq [-1/2, 1/2]\}\,.$$

The classical theorems of Beurling and Landau yields an almost complete characterization of sampling sets for the Paley-Wiener space.

Theorem 1. *(i) Beurling: If $D^-(\Lambda) > 1$ and Λ is separated, then Λ is a set of stable sampling for $\mathcal{B}$.*

(ii) Landau: If Λ is a set of stable sampling for $\mathcal{B}$, then $D^-(\Lambda) \geq 1$.

Item (i) is proved with complex analysis methods and Beurling's theory of weak limits of sets, (ii) uses the operator theory of localization operators.

The situation for general shift-invariant spaces is far from transparent. Necessary density conditions in the style of Landau were already proved early on: *Assume that* $\sum_{k\in\mathbb{Z}} \sup_{x\in[0,1]} |g(x+k)| < \infty$. *If* Λ *is set of stable sampling for* $V(g)$, *then* $D^-(\Lambda) \geq 1$.

Until recently sufficient conditions were formulated in terms of the maximum gap between consecutive samples and were mainly of qualitative nature, see e.g. the survey [1]. Let $\delta(\Lambda) = \sup_{j\in\mathbb{Z}}(\lambda_{j+1} - \lambda_j)$ maximum gap (or fill distance) and assume that Λ is (relatively) separated. Aldroubi and Feichtinger were the first to show that *for sufficiently small* $\delta(\Lambda)$, *depending on the generator* g, Λ *is a set of stable sampling for* $V(g)$. However, only in a few special cases, e.g. for B-splines, exponential B-splines, or totally positive functions of finite type, the best bound $\delta(\Lambda) < 1$ could be derived so far.

In this talk we report about recent progress for a general and natural class of generators, namely certain totally positive functions. Using Schoenberg's factorization of totally positive functions, we restrict ourselves to the subclass of totally positive functions whose Fourier transform factors as

$$\hat{g}(\xi) = ce^{-\gamma\xi^2} \prod_{j=1}^{N} (1 + 2\pi i\nu_j\xi)^{-1} \tag{1}$$

with $\gamma > 0$, $\nu_j \in \mathbb{R}$, $N \in \mathbb{N}$, which we call totally positive generators of Gaussian type. For such generators we were able to obtain a precise analogue of Beurling's result for Paley-Wiener space.

Theorem 2 ([3]). *Assume that* g *is totally positive of Gaussian type as in* (1). *If* $D^-(\Lambda) > 1$ *and* Λ *is separated, then* Λ *is a set of stable sampling.*

This theorem is optimal, since a sampling set satisfies always $D^-(\Lambda) \geq 1$.

The proof required a new combination of ideas from complex analysis (counting density of zeros), spectral invariance (off-diagonal decay of infinite matrix is preserved by inversion), and Beurling technique of weak limits of sets. By contrast we did not use the definition of total positivity.

As a modification we obtained a sampling theorem with derivatives. This problem occurs in contemporary data processing under the name event-based sampling or gradient-augmented sampling. For the formulation we need a multiplicity function $m : \Lambda \to \mathbb{N}$ such that $\sup_{\lambda\in\Lambda} m_\lambda < \infty$ which counts the number of derivatives at each sampling point of Λ. Correspondingly we use the weighted Beurling lower density

$$D^-(\Lambda, m_\Lambda) := \liminf_{r\to\infty} \inf_{x\in\mathbb{R}} \frac{1}{2r} \sum_{\lambda\in\Lambda\cap[x-r,x+r]} m_\lambda. \tag{2}$$

Then the corresponding result about sampling with derivatives is as follows:

Theorem 3 ([4]). *Let* g *be a totally positive function of Gaussian type as in* (1). *Let* $\Lambda \subseteq \mathbb{R}$ *be a separated set and let* m_Λ *be a bounded sequence of multiplicities. If*

$D^-(\Lambda, m_\Lambda) > 1$, then (Λ, m_Λ) is a sampling set for $V^2(g)$, i.e. for some constants $A, B > 0$,

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda} \sum_{k=0}^{m_\lambda - 1} |f^{(k)}(\lambda)|^2 \leq B\|f\|_2^2 \qquad \text{for all } f \in V(g)\,. \tag{3}$$

Again the result is optimal, because one can show that for a set satisfying (3) one has necessarily $D^-(\Lambda, m_\Lambda) \geq 1$.

A subtle connection of sampling in shift-invariant spaces and Gabor frames permits to translate the sampling theorems into new statement about Gabor frames. For this we denote the set of time-frequency shifts $\mathcal{G}(g, \Lambda \times \mathbb{Z}) = \{e^{2\pi ik\cdot} g(\cdot - \lambda) : k \in \mathbb{Z}, \lambda \in \Lambda\}$ and ask when this set is a frame (aka Gabor frame) for $L^2(\mathbb{R})$. The following result adds to the long history of Gabor frames [2] and to a conjecture of Ingrid Daubechies about the role of positivity in the theory of Gabor frames.

Theorem 4. *Assume that g is a totally positive function of Gaussian type. Let $\Lambda \subseteq \mathbb{R}$ be a separated set.*

Then $\mathcal{G}(g, \Lambda \times \mathbb{Z})$ is a Gabor frame for $L^2(\mathbb{R})$, if and only if $D^-(\Lambda) > 1$.

We formulate the case of rectangular lattices explicitly as a corollary.

Corollary 1. *Assume that g is a totally positive function of Gaussian type. Then $\mathcal{G}(g, \alpha\mathbb{Z} \times \beta\mathbb{Z})$ is a frame, if and only if $\alpha\beta < 1$.*

References

[1] A. Aldroubi and K. Gröchenig. Nonuniform sampling and reconstruction in shift-invariant spaces. *SIAM Rev.*, 43(4):585–620, 2001.

[2] K. Gröchenig. The mystery of Gabor frames. *J. Fourier Anal. Appl.*, 20(4):865–895, 2014.

[3] K. Gröchenig, J.-L. Romero and J. Stöckler. Sampling Theorems for Shift-invariant Spaces, Gabor Frames, and Totally Positive Functions *Invent. Math.* 211 (3), 1119 - 1148.

[4] K. Gröchenig, J.-L. Romero and J. Stöckler. Sharp results on sampling with derivatives in shift-invariant spaces and multi-window Gabor Frames. ArXiv https://arxiv.org/pdf/1712.07899.pdf

Synthesis of Logical Clifford Operators via Symplectic Geometry

Robert Calderbank

(joint work with Narayanan Rengaswamy, Swanand Kadhe, Henry Pfister)

Quantum error-correcting codes can be used to protect qubits involved in quantum computation. This requires that logical operators acting on protected qubits be translated to physical operators (circuits) acting on physical quantum states. We propose a mathematical framework for synthesizing physical circuits that implement logical Clifford operators for stabilizer codes. Circuit synthesis is enabled by representing the desired physical Clifford operator in $\mathbb{C}^{N\times N}$ as a partial $2m \times 2m$ binary symplectic matrix, where $N = 2^m$. We state and prove two theorems that use symplectic transvections to efficiently enumerate all binary symplectic matrices that satisfy a system of linear equations. As an important corollary of these

results, we prove that for an $[\![m, m-k]\!]$ stabilizer code every logical Clifford operator has $2^{k(k+1)/2}$ symplectic solutions. The desired physical circuits are then obtained by decomposing each solution as a product of elementary symplectic matrices, each corresponding to an elementary circuit. Our assembly of the possible physical realizations enables optimization over the ensemble with respect to a suitable metric. Furthermore, we show that any circuit that normalizes the stabilizer of the code can be transformed into a circuit that centralizes the stabilizer, while realizing the same logical operation. However, the optimal circuit for a given metric may not correspond to a centralizing solution. Our method of circuit synthesis can be applied to any stabilizer code, and this paper provides a proof of concept synthesis of universal Clifford gates for the $[\![6, 4, 2]\!]$ CSS code. We conclude with a classical coding-theoretic perspective for constructing logical Pauli operators for CSS codes. Since our circuit synthesis algorithm builds on the logical Pauli operators for the code, this paper provides a complete framework for constructing all logical Clifford operators for CSS codes. Programs implementing the algorithms in this paper, which includes routines to solve for binary symplectic solutions of general linear systems and our overall circuit synthesis algorithm, can be found at `https://github.com/nrenga/symplectic-arxiv18a`.

References

[1] N. Rengaswamy, R. Calderbank, S. Kadhe, and H. Pfister, "Synthesis of logical clifford operators via symplectic geometry," in *Proc. IEEE Int. Symp. Inform. Theory*, 2018.

[2] N. Rengaswamy, R. Calderbank, S. Kadhe, and H. D. Pfister, "Synthesis of Logical Clifford Operators via Symplectic Geometry," *arXiv preprint arXiv:1803.06987*, 2018.

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

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

[5] D. Gottesman, "A Theory of Fault-Tolerant Quantum Computation," *arXiv preprint arXiv:quant-ph/9702029*, 1997. [Online]. Available: http://arxiv.org/pdf/quant-ph/9702029.pdf.

[6] A. Calderbank, E. Rains, P. Shor, and N. Sloane, "Quantum error correction via codes over GF(4)," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1369–1387, Jul 1998.

[7] M. M. Wilde, "Logical operators of quantum codes," *Phys. Rev. A*, vol. 79, no. 6, p. 062322, 2009.

[8] D. Gottesman, "An Introduction to Quantum Error Correction and Fault-Tolerant Quantum Computation," *arXiv preprint arXiv:0904.2557*, 2009. [Online]. Available: http://arxiv.org/pdf/0904.2557.pdf.

[9] M. Grassl and M. Roetteler, "Leveraging automorphisms of quantum codes for fault-tolerant quantum computation," in *Proc. IEEE Int. Symp. Inform. Theory*, pp. 534–538, IEEE, Jul 2013.

[10] R. Chao and B. W. Reichardt, "Fault-tolerant quantum computation with few qubits," *arXiv preprint arXiv:1705.05365*, 2017. [Online]. Available: http://arxiv.org/pdf/1705.05365.pdf.

Learning via nonconvex optimization: ReLUs, neural nets, and submodular maximization

Mahdi Soltanolkotabi

Many problems of contemporary interest in signal processing and machine learning involve highly non-convex optimization problems. While nonconvex problems are known to be intractable in general, simple local search heuristics such as (stochastic) gradient descent are often surprisingly effective at finding global/high quality optima on real or randomly generated data. In this note we summarize some recent results explaining the success of these heuristics focusing on two problems: (1) learning the optimal weights of the shallowest of neural networks consisting of a single Rectified Linear Unit (ReLU), (2) learning over-parameterized neural networks with a single hidden layer. In the talk we also discussed a third problem of maximizing submodular functions (we omit this description here due to space limitation and refer to [3] for detail on this problem). This summary is based on our papers [1, 2, 3]. We refer to these papers for a comprehensive discussion on related work in these areas.

1. Problem I: Learning ReLUs

Nonlinear data-fitting problems are fundamental to many supervised learning tasks in signal processing and machine learning. Given training data consisting of n pairs of input features $\boldsymbol{x}_i \in \mathbb{R}^d$ and desired outputs $\boldsymbol{y}_i \in \mathbb{R}$ we wish to infer a function that best explains the training data. One form of nonlinearity which is of particular interest in modern learning is that of fitting Rectified Linear Units (ReLUs) to the data which are functions $\phi_{\boldsymbol{w}} : \mathbb{R}^d \to \mathbb{R}$ of the form

$$\phi_{\boldsymbol{w}}(\boldsymbol{x}) = \max\left(0, \langle \boldsymbol{w}, \boldsymbol{x} \rangle\right).$$

A natural approach to fitting ReLUs is via nonlinear least-squares of the form

$$(1) \qquad \min_{\boldsymbol{w} \in \mathbb{R}^d} \quad \mathcal{L}(\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^{n} \left(\max\left(0, \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle\right) - y_i\right)^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{w}) \leq R,$$

with $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ a regularization function that encodes prior information on the weight vector.

A simple heuristic for optimizing (1) is to use projected gradient descent like updates. A-priori it is completely unclear why such local search heuristics should converge for problems of the form (1), as not only the regularization function maybe nonconvex but also the loss function! Our result aims to explain why gradient descent is effective in this setting.

Theorem 1. *Let $\boldsymbol{w}^* \in \mathbb{R}^d$ be an arbitrary weight vector and $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ be a proper function (convex or nonconvex). Suppose the feature vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ are i.i.d. Gaussian random vectors distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ with the corresponding labels given by*

$$\boldsymbol{y}_i = \max\left(0, \langle \boldsymbol{x}_i, \boldsymbol{w}^* \rangle\right).$$

To estimate $\boldsymbol{w}^*$, we start from the initial point $\boldsymbol{w}_0 = \mathbf{0}$ and apply the Projected Gradient (PGD) updates of the form

$$\boldsymbol{w}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{w}_\tau - \mu_\tau \nabla \mathcal{L}(\boldsymbol{w}_\tau)\right), \tag{2}$$

with $\mathcal{K} := \{\boldsymbol{w} \in \mathbb{R}^d : \ \mathcal{R}(\boldsymbol{w}) \leq \mathcal{R}(\boldsymbol{w}^*)\}$. Also set the learning parameter sequence $\mu_0 = 2$ and $\mu_\tau = 1$ for all $\tau = 1, 2, \ldots$. Also assume

$$n > cn_0, \tag{3}$$

holds for a fixed numerical constant c. Here, n_0 is a lower bound on the minimum number of samples required using any algorithm (see [1] for a precise definition). Then there is an event of probability at least $1 - 9e^{-\gamma n}$ such that on this event the updates (2) obey

$$\|\boldsymbol{w}_\tau - \boldsymbol{w}^*\|_{\ell_2} \leq \left(\frac{1}{2}\right)^\tau \|\boldsymbol{w}^*\|_{\ell_2}. \tag{4}$$

Here γ is a fixed numerical constant.

Despite the nonconvexity of both the objective and regularizer, the theorem above shows that with a near minimal number of data samples, projected gradient descent provably learns the original weight vector $\boldsymbol{w}^*$ without getting trapped in any local optima.

2. Problem II: Learning over-parameterized shallow neural nets

Neural network architectures (a.k.a. deep learning) have recently emerged as powerful tools for automatic knowledge extraction from raw data. These learning architectures have lead to major breakthroughs in many applications. Despite their wide empirical use the mathematical success of these architectures remains a mystery. The main challenge is that training neural networks correspond to extremely high-dimensional and nonconvex optimization problems and it is not clear how to provably solve them to global optimality. These networks are trained successfully in practice via local search heuristics on real or randomly generated data. In particular, over-parameterized neural networks-where the number of parameters exceed the number of data samples-can be optimized to global optimality using local search heuristics such as gradient or stochastic gradient methods. In our paper [2] we provide theoretical insights into this phenomenon by developing a better understanding of optimization landscape of such over-parameterized shallow neural networks.

We discuss the main results in [2]. The results we present here focuses on understanding the global landscape of neural network optimization with one hidden layer with quadratic activation functions. The paper [2] also contains results on the local convergence of gradient descent that applies to a broad set of activation functions. We omit these results do to space limitations.

Theorem 2. *Assume we have an arbitrary data set of input/label pairs $\boldsymbol{x}_i \in \mathbb{R}^d$ and y_i for $i = 1, 2, \ldots, n$. Consider a neural network of the form*

$$\boldsymbol{x} \mapsto \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}),$$

with $\phi(z) = z^2$ a quadratic activation and $\boldsymbol{W} \in \mathbb{R}^{k\times d}$, $\boldsymbol{v} \in \mathbb{R}^k$ denoting the weights connecting input to hidden and hidden to output layers. We assume $k \geq 2d$ and set the weights of the output layer $\boldsymbol{v}$ so as to have at least d positive entries and at least d negative entries. Then, the training loss as a function of the weights $\boldsymbol{W}$ of the hidden layer

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i)\right)^2,$$

obeys the following two properties.

- *There are no spurious local minima, i.e. all local minima are global.*
- *All saddle points have a direction of strictly negative curvature. That is, at a saddle point $\boldsymbol{W}_s$ there is a direction $\boldsymbol{U} \in \mathbb{R}^{k\times d}$ such that*

$$\mathrm{vect}(\boldsymbol{U})^T \nabla^2 \mathcal{L}(\boldsymbol{W}_s) \mathrm{vect}(\boldsymbol{U}) < 0.$$

Furthermore, for almost every data inputs $\{\boldsymbol{x}_i\}_{i=1}^n$, as long as

$$d \leq n \leq cd^2,$$

the global optimum of $\mathcal{L}(\boldsymbol{W})$ is zero. Here, $c > 0$ is a fixed numerical constant.

The above result states that given an arbitrary data set, the optimization landscape of fitting neural networks have favorable properties that facilitate finding globally optimal models. In particular, by setting the weights of the last layer to have diverse signs all local minima are global minima and all saddles have a direction of negative curvature. This in turn implies that gradient descent on the input-to-hidden weights, when initialized at random, converges to a global optima. All of this holds as long as the neural network is sufficiently wide in the sense that the number of hidden units exceed the dimension of the inputs by a factor of two ($k \geq 2d$).

References

[1] M. Soltanolkotabi, *Learning ReLUs via gradient descent*, Neural Information Processing Systems (NIPS 2017).

[2] Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Under revision in IEEE Trans. on Info. Theory

[3] H. Hassani, M. Soltanolkotabi and A. Karbasi., *Gradient methods for submodular maximization*, Neural Information Processing Systems (NIPS 2017).

SUNLayer: stable denoising with generative networks

Soledad Villar
(joint work with Dustin G. Mixon)

Exploiting the structure of signals is a fundamental idea in signal processing. For instance, natural images are sparse on wavelets basis [4], and sparsity allows the recovery of signals from few measurements (à la compressed sensing [3]).

The current trend, that takes advantage of the empirical success of deep learning, is to learn the structure of the signal first, and then exploit it. One way to represent the structure of signals is through a generative model. Informally, a generative model can be thought as a form of parametrization of the data

$$G : \mathbb{R}^n \to \mathbb{R}^N \quad n \ll N$$

such that $G(\mathbb{R}^n)$ is a proxy for the probability density of the data of interest.

Very impressive generative models have been produced (see for instance [1]) using autoencoders and generative adversarial networks [6]. However, there does not seem to currently exist a provable way to produce generative models successfully and even when the generative model produced is useful for application purposes it is not clear whether a reasonable data distribution is actually learned. Producing generative models that are useful for applications is a very active research area within the machine learning community.

1. Inverse problems with generative models

If we have a good generative model we can do amazing things with it. For instance, recent work by Bora, Jalal, Price and Dimakis [2] empirically shows that a generative model obtained from a generative adversarial network can be used to solve the compressed sensing problem with 10 times fewer measurements than classical compressed sensing requires. The key idea is to replace the sparse signal assumption by assuming the signal is close to the range of the generative model G. Their theoretical result shows that under mild hypothesis, if $y = Ax^* + \eta$ (η is the noise), then

$$z^* = \arg\min_z \|AG(z) - y\|_2 \tag{1}$$

satisfies $G(z^*) \approx x^*$ (see Theorem 1.1 of [2]).

However, it is not obvious that one can efficiently solve the optimization problem (1) since its landscape may a priori have many local minima. Recent work by Hand and Voroninski [7] shows the optimization problem (1) can be solved using local methods provided that

$$G = (\rho \circ G_\ell) \circ \ldots \circ (\rho \circ G_1)$$

where $\rho(t) = \mathrm{ReLU}(t) = \max\{0, t\}$ and G_i are matrices with i.i.d. Gaussian entries properly scaled. There is no learning in this setting.

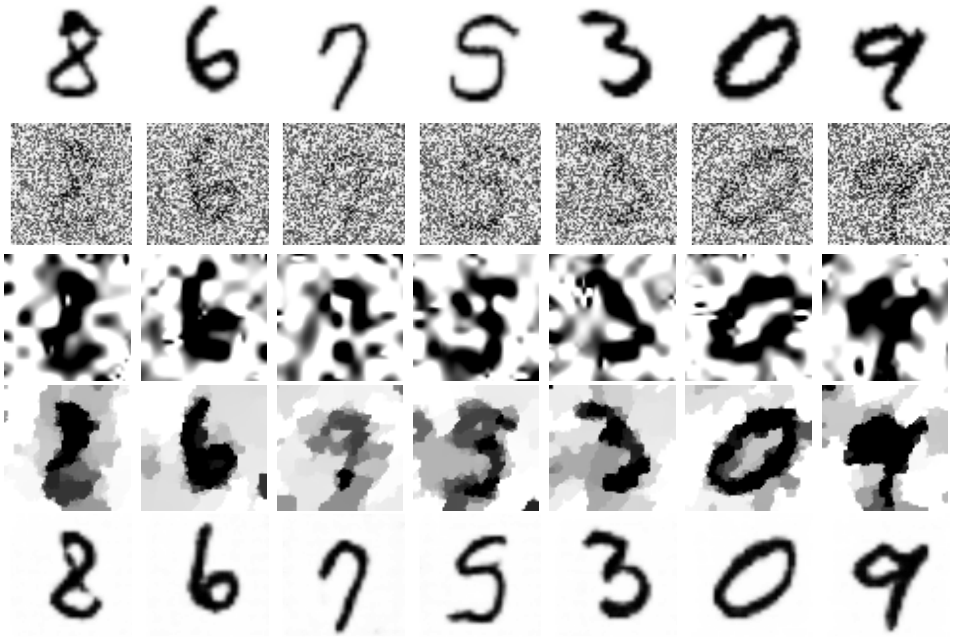

FIGURE 1. Denoising with generative priors
(First line) Digits from the MNIST test set ([8]). **(Second line)** random noise is added to the digits. **(Third line)** Denoising of images by shrinkage in wavelet domain ([5]). **(Fourth line)** Denoising by minimizing total variation ([11]). **(Fifth line)** We train a GAN using the training set of MNIST to obtain a generative model G. We denoise by finding the closest element in the image of G using stochastic grading descent.

2. SUNLayer and denoising with generative models

Motivated by both works [2, 7], in our paper [9] we study the simpler inverse problem of signal denoising with generative networks. The aim of [9] is to explain the phenomenon illustrated in Figure 1, i.e. given $y = G(x^*) + \eta$ a noisy signal then one can denoise by finding

$$z^* = \arg\min_z \|G(z) - y\|_2 \tag{2}$$

and the optimization problem (2) can be solved by local methods like gradient descent (i.e. (2) has no spurious critical points).

We consider a simpler model for a generative model inspired by neural networks. One layer of the SUNLayer (spherical uniform neural layer) is defined as

$$\begin{aligned} L_n : S^n &\to \mathscr{L}^2(S^n) \\ x &\mapsto \rho(\langle x, \cdot \rangle), \end{aligned}$$

where ρ is an arbitrary activation function. We aim to answer what properties of activation functions allow denoising with local methods in this simplified model.

Consider the decomposition of ρ in the Gegenbauer polynomials (choosing the correct normalization will be important). If $\rho(t) = \sum_{k=0}^{\infty} a_k \varphi_{k,n}(t)$, we define $g_\rho(t) = \sum_{k=0}^{\infty} a_k^2 \varphi_{k,n}(t)$. Then our main result shows that the critical points of (2) are close to $\pm x^*$ if $\inf_{t\in[-1,1]} |g'_\rho(t)|$ is not too small in comparison with $\|\eta\|$.

3. Spherical harmonics

Squaring the coefficients in the Gegenbauer decomposition may look a priori mysterious. However, it shows up naturally due to the nice properties of the spherical harmonics. Consider the simplified setting when there is no noise. We have

$$\arg\min_z \|L_n(z) - L(x^*)\|^2 = \arg\min_z \|L_n(z)\|^2 + \|L_n(x^*)\|^2 - 2\langle L_n(z), L_n(x^*)\rangle$$

and $\|L_n(x)\|^2 = \int_{S^n} \rho(\langle x, y\rangle)^2 dy = c_{\rho,n}$ independent of x. Now we use the relationship between the Gegenbauer polynomials and the spherical harmonics (see chapter 2 of [10]). Decompose $\mathscr{L}^2(S^n) = \oplus_{k=0}^{\infty} \mathcal{H}_n^k(S^n)$ where $\mathcal{H}_n^k(S^n)$ are the spherical harmonics (homogeneous polynomials in $n+1$ variables, of degree k, with Laplacian 0, restricted to the sphere). In particular $\mathcal{H}_n^k(S^n)$ is a finite dimensional vector space. Let $\{Y_1, \dots Y_r\}$ a basis of $\mathcal{H}_n^k(S^n)$, then define the bilinear form $F_k(\sigma, \tau) = \sum_{s=1}^{r} Y_s(\sigma)\overline{Y_s(\tau)}$ for $\sigma, \tau \in S^n$.

It turns out F_k is a reproducing kernel: $\langle F_k(x,\cdot), F_k(y,\cdot)\rangle = F_k(x,y)$ and it also satisfies that $F_k(x,y)$ only depends on the inner product $t := \langle x, y\rangle$ and in fact $F_k(x,y) = \varphi_{k,n}(t)$. Then

$$\begin{aligned}
\langle L_n(z), L_n(x^*)\rangle &= \langle \rho(\langle z,\cdot\rangle), \rho(\langle x^*,\cdot\rangle)\rangle \\
&= \langle \sum_{k=0}^{\infty} a_k \varphi_{k,n}(\langle z,\cdot\rangle), \sum_{k=0}^{\infty} a_k \varphi_{k,n}(\langle x^*,\cdot\rangle)\rangle \\
&= \langle \sum_{k=0}^{\infty} a_k F_k(z,\cdot), \sum_{k=0}^{\infty} a_k F_k(\langle x^*,\cdot\rangle\rangle \\
&= \sum_{k=0}^{\infty} a_k^2 \langle F_k(z,\cdot), F_k(\langle x^*,\cdot\rangle\rangle \\
&= \sum_{k=0}^{\infty} a_k^2 \varphi_{n,k}(\langle z, x^*\rangle).
\end{aligned}$$

Therefore

$$\arg\min_z \|L_n(z) - L(x^*)\|^2 = \arg\max_z \langle L_n(z), L_n(x^*)\rangle = \arg\max_z \sum_{k=0}^{\infty} a_k^2 \varphi_{n,k}(\langle z, x^*\rangle)$$

and a simple computation shows that the only critical points are $z = \pm x^*$ if $g'_\rho(t) > 0$ for all $t \in [-1,1]$. The analysis for the noisy case we do in [9] is still simple but more interesting since it involves considering tight frames in $\mathcal{H}_n^k(S^n)$.

References

[1] Berthelot, David, Tom Schumm, and Luke Metz. *Began: Boundary equilibrium generative adversarial networks.* arXiv preprint arXiv:1703.10717 (2017).

[2] Bora, Ashish, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. *Compressed sensing using generative models.* arXiv preprint arXiv:1703.03208 (2017).

[3] Candes, Emmanuel J., Justin K. Romberg, and Terence Tao. *Stable signal recovery from incomplete and inaccurate measurements.* Communications on pure and applied mathematics 59, no. 8 (2006): 1207-1223.

[4] Daubechies, Ingrid. *Ten lectures on wavelets.* Vol. 61. Siam, 1992.

[5] Donoho, David L., and Jain M. Johnstone. *Ideal spatial adaptation by wavelet shrinkage.* Biometrika 81, no. 3 (1994): 425-455.

[6] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. *Generative adversarial nets.* In Advances in neural information processing systems, (2014): 2672-2680.

[7] Hand, Paul, and Vladislav Voroninski. *Global Guarantees for Enforcing Deep Generative Priors by Empirical Risk.* arXiv preprint arXiv:1705.07576 (2017).

[8] LeCun, Yann. *The MNIST database of handwritten digits.* http://yann. lecun. com/exdb/mnist/ (1998).

[9] Mixon, Dustin G., and Soledad Villar. *SUNLayer: Stable denoising with generative networks.* arXiv preprint arXiv:1803.09319 (2018).

[10] Morimoto, Mitsuo. *Analytic functionals on the sphere and their Fourier-Borel transformations.* Complex Analysis Banach Center Publications 11 (1983).

[11] Rudin, Leonid I., Stanley Osher, and Emad Fatemi. *Nonlinear total variation based noise removal algorithms.* Physica D: nonlinear phenomena 60, no. 1-4 (1992): 259-268.

Random Initialization in Nonconvex Phase Retrieval

Yuxin Chen

(joint work with Cong Ma, Yuejie Chi, Jianqing Fan)

Suppose we are interested in learning an unknown object $\mathbf{x}^\natural \in \mathbb{R}^n$, but only have access to a few quadratic equations of the form

$$y_i = \left(\mathbf{a}_i^\top \mathbf{x}^\natural\right)^2, \qquad 1 \le i \le m, \tag{1}$$

where y_i is the sample we collect and $\mathbf{a}_i$ is the design vector known *a priori*. Is it feasible to reconstruct $\mathbf{x}^\natural$ in an accurate and efficient manner?

The problem of solving systems of quadratic equations (1) spans multiple domains including physical sciences and machine learning. A natural strategy for inverting the system of quadratic equations (1) is to solve the following nonconvex least squares estimation problem

$$\text{minimize}_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) := \frac{1}{4m}\sum_{i=1}^{m}\left[\left(\mathbf{a}_i^\top \mathbf{x}\right)^2 - y_i\right]^2. \tag{2}$$

Under Gaussian designs where $\mathbf{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, the solution to (2) is known to be exact — up to some global sign — with high probability, as soon as the number m of equations (samples) exceeds the order of the number n of unknowns. However, the loss function in (2) is highly nonconvex, thus resulting in severe computational challenges. Fortunately, in spite of nonconvexity, a variety of optimization-based

methods are shown to be effective in the presence of proper statistical models. Arguably, one of the simplest algorithms for solving (2) is vanilla gradient descent (GD), which attempts recovery via the update rule

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \nabla f\left(\mathbf{x}^t\right), \qquad t = 0, 1, \cdots \tag{3}$$

with η_t being the stepsize / learning rate. The above iterative procedure is also dubbed *Wirtinger flow* for phase retrieval, which can accommodate the complex-valued case as well. This simple algorithm is remarkably efficient under Gaussian designs: in conjunction with carefully-designed initialization and stepsize rules, GD provably converges to the truth $\mathbf{x}^\natural$ at a linear rate[1], provided that the ratio m/n of the number of equations to the number of unknowns exceeds some logarithmic factor.

One crucial element in prior convergence analysis is initialization. In order to guarantee linear convergence, prior works typically recommend spectral initialization or its variants. Two important features are worth emphasizing:

- $\mathbf{x}^0$ falls within a local ℓ_2-ball surrounding $\mathbf{x}^\natural$ with a reasonably small radius, where $f(\cdot)$ enjoys strong convexity;
- $\mathbf{x}^0$ is incoherent with all the design vectors $\{\mathbf{a}_i\}$ — in the sense that $|\mathbf{a}_i^\top \mathbf{x}^0|$ is reasonably small for all $1 \le i \le m$ — and hence $\mathbf{x}^0$ falls within a region where $f(\cdot)$ enjoys desired smoothness conditions.

These two properties taken collectively allow gradient descent to converge rapidly from the very beginning.

The enormous success of spectral initialization gives rise to a curious question: is carefully-designed initialization necessary for achieving fast convergence? A strategy that practitioners often like to employ is to initialize GD randomly. The advantage is clear: compared with spectral methods, random initialization is model-agnostic and is usually more robust vis-a-vis model mismatch. Despite its wide use in practice, however, GD with random initialization is poorly understood in theory.

In this work, we prove that under Gaussian designs (i.e. $\mathbf{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$), gradient descent — when randomly initialized — yields an ϵ-accurate solution in $O\big(\log n + \log(1/\epsilon)\big)$ iterations given nearly minimal samples (up to some logarithmic factor), thus achieving near-optimal computational and sample complexities at once. This provides the first global convergence guarantee concerning vanilla gradient descent for phase retrieval, without the need of (i) carefully-designed initialization, (ii) sample splitting, or (iii) sophisticated saddle-point escaping schemes. All of these are achieved by exploiting the statistical models in analyzing optimization algorithms, via a leave-one-out approach that enables the decoupling of certain statistical dependency between the gradient descent iterates and the data.

[1]An iterative algorithm is said to enjoy linear convergence if the iterates $\{\mathbf{x}^t\}$ converge geometrically fast to the minimizer $\mathbf{x}^\natural$.

Optimal weighted least squares approximations

Albert Cohen
(joint work with Benjamin Arras, Markus Bachmayr and Giovanni Migliorati)

We consider the problem of approximating an unknown function $u \in L^2(X,\rho)$ from its evaluation at given sampling points $x^1,\dots,x^n \in X$, where $X \subset \mathbb{R}^d$ is a general domain and ρ a probability measure. The approximation $\tilde{u}$ is picked in a linear space V_m where $m = \dim(V_m)$. We measure accuracy in the Hilbertian norm

$$\|v\| = \left(\int_X |v(x)|^2 d\rho\right)^{1/2} = \|v\|_{L^2(X,\rho)},$$

where ρ is a probability measure over X. The error of best approximation is defined by

$$e_m(u) := \min_{v\in V_m} \|u-v\|,$$

The method is said to be near-optimal (or instance optimal with constant C) if the comparison

$$\|u-\tilde{u}\| \le C e_m(u),$$

holds for all u, where $C>1$ is some fixed constant.

For a given probability measure ρ and approximation space V_m of interest, a relevant question is whether instance optimality can be achieved with sample size n that is moderate, ideally linear in m. Recent results of [3, 5] for polynomial spaces and [2] in a general approximation setting, show that this objective can be achieved by certain random sampling schemes in the general framework of *weighted least squares* methods. The approximation $\tilde{u}$ is defined as the solution to

$$\min_{v\in V_m} \frac{1}{n}\sum_{i=1}^n w(x^i)|y^i - v(x^i)|^2,$$

where w is a positive function and the x^i are independently drawn according to a probability measure μ, that satisfy the constraint

$$w\, d\mu = d\rho.$$

The case $w=1$ and $\mu=\rho$ corresponds to the standard unweighted least squares method.

We denote by $\|\cdot\|_n$ the discrete Euclidean norm defined by

$$\|v\|_n^2 := \frac{1}{n}\sum_{i=1}^n w(x^i)|v(x^i)|^2,$$

and by $\langle\cdot,\cdot\rangle_n$ the associated inner product. The solution $\tilde{u}$ may be thought of as an orthogonal projection of u onto V_m for this norm. Expanding it into

$$\tilde{u} = \sum_{j=1}^m c_j \varphi_j,$$

in a basis $\{\varphi_1, \ldots, \varphi_m\}$ of V_m, the coefficient vector $\mathbf{c} = (c_1, \ldots, c_m)^T$ is solution to the linear system

$$\mathbf{G}\mathbf{c} = \mathbf{d},$$

where $\mathbf{G}$ is the Gramian matrix for the inner product $\langle \cdot, \cdot \rangle_n$ with entries

$$\mathbf{G}_{j,k} := \langle \varphi_j, \varphi_k \rangle_n = \frac{1}{n} \sum_{i=1}^{n} w(x^i) \varphi_j(x^i) \varphi_k(x^i),$$

and the vector $\mathbf{d}$ has entries $\mathbf{d}_k = \frac{1}{n} \sum_{i=1}^{n} y^i \varphi_k(x^i)$. The solution $\mathbf{c}$ always exists and is unique when $\mathbf{G}$ is invertible. When $\{\varphi_1, \ldots, \varphi_m\}$ is an $L^2(X, \rho)$-orthonormal basis of V_m, one has

$$\mathbb{E}(\mathbf{G}) = \mathbf{I}.$$

The stability and accuracy analysis of the weighted least squares method is related to the amount of deviation between $\mathbf{G}$ and its expectation $\mathbf{I}$ measured in the spectral norm. This deviation also describe the closeness of the norms $\|\cdot\|$ and $\|\cdot\|_n$ over the space V_m, since one has

$$\|\mathbf{G} - \mathbf{I}\| \leq \delta \quad \Longleftrightarrow \quad (1-\delta)\|v\|^2 \leq \|v\|_n^2 \leq (1+\delta)\|v\|^2, \quad v \in V_m.$$

The choice of a sampling measure μ that differs from the error norm measure ρ appears to be critical in order to obtain stable and accurate approximations with an optimal sampling budget. The optimal sampling measure and weights are given by

$$\mu_m = \frac{k_m}{m} d\rho \quad \text{and} \quad w_m = \frac{m}{k_m},$$

where k_m is the so-called Christoffel function defined by

$$k_m(x) = \sum_{j=1}^{m} |\varphi_j(x)|^2,$$

with $\{\varphi_1, \ldots, \varphi_m\}$ any $L^2(X, \rho)$-orthonormal basis of V_m. With such choices, the following result can be established, see [2, 1].

Theorem 1. *With the above choice μ_m of sampling measure, for any $0 < \varepsilon < 1$, the condition*

$$n \geq cm(\ln(2m) - \ln(\varepsilon)), \quad c := \gamma^{-1} = \frac{2}{1 - \ln 2},$$

implies the following stability and instance optimality properties:

$$\Pr\Big(\|\mathbf{G} - \mathbf{I}\| \geq \frac{1}{2}\Big) \leq \varepsilon.$$

and

$$\mathbb{E}(\|u - \tilde{u}\|^2) \leq \Big(1 + \frac{c}{\ln(2m) - \ln(\varepsilon)}\Big) e_m(u)^2 + \varepsilon \|u\|^2.$$

In summary, when using the optimal sampling measure μ_m, stability and instance optimality can be achieved in the near linear regime

$$n = n(m) = n_\varepsilon(m) := \lceil c\, m\, (\ln(2m) - \ln \varepsilon) \rceil,$$

where ε controls the probability of failure.

In various practical applications, the space V_m is picked within a family $(V_m)_{m\geq 1}$ that has the nestedness property

$$V_1 \subset V_2 \subset \cdots$$

and accuracy is improved by raising the dimension m. The sequence $(V_m)_{m\geq 1}$ may either be a priori defined, or adaptively generated, which means that the way V_m is refined into V_{m+1} may depend on the result of the least squares computation. In this setting, we are facing the difficulty that the optimal measure μ_m varies with m.

In order to maintain an optimal sampling budget, one should avoid the option of drawing a new sample $S_m = \{x_m^1, \ldots, x_m^n\}$ of increasing size $n = n(m)$ at each step m. For this purpose, we observe that the optimal measure μ_m enjoys the mixture property

$$\mu_{m+1} = \Big(1 - \frac{1}{m+1}\Big)\mu_m + \frac{1}{m+1}\sigma_{m+1}, \quad \text{where} \quad d\sigma_m := |\varphi_m|^2 d\rho.$$

As noticed in [4], this leads naturally to sequencial sampling strategies where the sample S_m is recycled for generating S_{m+1}. Here is one instance of such a strategy that was studied in [1].

Algorithm 1 Sequential sampling

input: sample S_m from μ_m
output: sample S_{m+1} from μ_{m+1}

for $i = 1, \ldots, n(m)$ **do**
  draw a_i uniformly distributed in $\{1, \ldots, m+1\}$
  if $a_i = m+1$ **then**
    draw x_{m+1}^i from σ_{m+1}
  else
    set $x_{m+1}^i := x_m^i$
  end if
end for
for $i = n(m)+1, \ldots, n(m+1)$ **do**
  draw x_{m+1}^i from μ_{m+1}.
end for

The interest of this sequencial sampling strategy is that the total number of sample C_m which has been generated after m steps remains within the same order as the near optimal budget $n(m)$. More precisely, the following result is established in [1].

Theorem 2. *For Algorithm 1, one has*

$$\mathbb{E}(C_m) \leq n(m) + n(m-1) + 1,$$

and for any $\tau \in [0,1]$,

$$\Pr\big(C_m \geq n(m) + (1+\tau)(n(m-1)+1)\big) \leq M_\tau e^{-\frac{\tau^2}{6}n(m-1)}$$

with $M_\tau := e^{\frac{2c\tau^2}{3}}$.

It should be noted that these results are completely independent of the choice of the spaces $(V_m)_{m\geq 1}$, as well as of the spatial dimension d of the domain X. One natural perspective is to develop adaptive least square methods in various context (wavelet refinements, high dimensional sparse polynomials) based on such sequencial sampling strategies.

References

[1] B. Arras, M. Bachmayr and A. Cohen, *Sequential sampling for optimal weighted least squares approximations in hierarchical spaces*, preprint (2018).

[2] A. Cohen and G. Migliorati, *Optimal weighted least squares methods,* SMAI Journal of Computational Mathematics **3** (2017), 181–203.

[3] A. Doostan and J. Hampton, *Coherence motivated sampling and convergence analysis of least squares polynomial Chaos regression*, Computer Methods in Applied Mechanics and Engineering **290** (2015), 73-97.

[4] A. Doostan and J. Hampton, *Basis Adaptive Sample Efficient Polynomial Chaos (BASE-PC)*, preprint (2017), arXiv:1702.01185.

[5] J.D. Jakeman, A. Narayan, and T. Zhou, *A Christoffel function weighted least squares algorithm for collocation approximations*, preprint (2016), arXiv:1412.4305.

On the Expressive Power of Deep Learning: A Tensor Analysis

Nadav Cohen

(joint work with Or Sharir, Yoav Levine, Ronen Tamari, David Yakira, Amnon Shashua)

It is widely accepted that the driving force behind convolutional networks, and deep learning in general, is the expressive power that comes with depth, i.e. the ability to compactly represent rich and effective spaces of functions through compositionality. Despite the vast empirical evidence supporting this belief, formal arguments to date are scarce. In particular, the machine learning community lacks satisfactory analyses of *depth efficiency*, a concept which refers to a situation where a deep network of polynomial size realizes a function that cannot be realized (or approximated) by a shallow network unless the latter has super-polynomial size. Moreover, even with a concrete understanding of depth efficiency, the mystery behind the expressive power of deep learning would still remain. Deep networks of polynomial size realize a small fraction of all possible functions, thus even if depth efficiency holds almost always, meaning the space of functions efficiently realizable by deep networks is much larger than that efficiently realizable by shallow networks, it still does not explain why deep networks are effective in practice. To

address this question one must consider the *inductive bias*, i.e. the assumptions regarding functions required for real-world tasks that are implicitly encoded into deep networks. In the series of papers described hereafter, we derive an equivalence between convolutional networks and *tensor decompositions*, and use it to analyze, for the first time, the depth efficiency and inductive bias of convolutional networks.

We begin in [3] by constructing a universal hypotheses space over instances defined as tuples of vectors, which in the context of images, corresponds to representation via local patches. The hypotheses space is constructed as a tensor product of finite-dimensional function spaces over the local structures. A general hypothesis may thus be expressed as a linear combination over an exponentially large basis of product functions, where the coefficients of the linear combination are naturally viewed as a high-order tensor (every mode in the tensor corresponds to a patch in the input). Naive computation of hypotheses is intractable, but by applying *hierarchical tensor decompositions* to coefficient tensors, efficient computation becomes possible. Moreover, circuits realizing the computations form a special case of convolutional networks. Namely, they are convolutional networks with linear activation and product pooling, and we accordingly refer to them as *convolutional arithmetic circuits*. The key observation is that there is a one-to-one correspondence between the type of decomposition applied to a coefficient tensor, and the structure of the convolutional arithmetic circuit computing the hypothesis (number of hidden layers, number of channels in each hidden layer, sizes and shapes of pooling windows etc.). This facilitates the study of networks through analysis of their corresponding tensor decompositions, bringing forth a plurality of mathematical tools from domains such as matrix algebra and measure theory.

We show that classic *CANDECOMP/PARAFAC (CP) decomposition* corresponds to a shallow network with global pooling in its single hidden layer. The recently introduced *Hierarchical Tucker (HT) decomposition* corresponds to a deep network with multiple hidden layers, where the sizes of pooling windows (and the resulting network depth) depend on the structure of the mode tree underlying the decomposition. By analyzing tensors generated by CP and HT decompositions in terms of their ranks when subject to canonical matrix arrangements, we show that besides a set of Lebesgue measure zero, all weight settings for a deep network lead to depth efficient functions. That is to say, besides a negligible set, all functions realizable by a deep network of polynomial size cannot be realized (or approximated) by a shallow network unless the latter has super-polynomial size. Such result, which we refer to as *complete depth efficiency*, has never before been established for any deep learning architecture, convolutional networks in particular.

Convolutional arithmetic circuits comprise the fundamental ingredients of convolutional networks – locality, weight sharing and pooling. We have implemented and evaluated such circuits (a.k.a. *SimNets*), showing that they deliver promising results on various visual recognition benchmarks [1, 2]. Nonetheless, they have yet

to reach widespread use, especially compared to *convolutional rectifier networks* – the most successful variant of convolutional networks to date. Convolutional rectifier networks are characterized by ReLU activation and max or average pooling. They do not posses the algebraic nature of convolutional arithmetic circuits, thus it is unclear to what extent our results from [3] apply to such networks.

To facilitate an analysis of convolutional rectifier networks, we head on in [4] and define *generalized tensor decompositions* as constructs that are obtained by replacing the multiplication operator in standard tensor decompositions with a general associative and commutative operator $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Given convolutional networks with activation σ (e.g. $\sigma(z) = \max\{z, 0\}$ for ReLU) and pooling P (e.g. $P\{c_i\}_i = max\{c_i\}_i$ for max), we define the *activation-pooling operator* $g_{\sigma/P}(a, b) := P\{\sigma(a), \sigma(b)\}$. Apparently, if $g_{\sigma/P}$ is associative and commutative, the generalized tensor decompositions it gives rise to are equivalent to convolutional networks with activation σ and pooling P, where again, there is a one-to-one correspondence between the type of a decomposition and the structure of its respective network. With convolutional rectifier networks the activation-pooling operator $g_{\sigma/P}$ is indeed associative and commutative, thus the equivalence holds. We make use of it to analyze the expressive properties of such networks, and, surprisingly, find that in contrast to convolutional arithmetic circuits, with convolutional rectifier networks depth efficiency is not complete. There are still functions efficiently realizable by deep networks and not by shallow ones, but these are not as common – the set of functions in a deep network's hypotheses space which can be realized (or approximated) by polynomially-sized shallow networks is non-negligible (has positive Lebesgue measure in the deep network's weight space). We interpret this result as indicating that in terms of expressiveness, the popular convolutional rectifier networks are inferior to the recently introduced convolutional arithmetic circuits. Of course, to take advantage of a machine learning model, it is not enough for it to be expressive, we have to be able to effectively train it as well. Over the years, a huge body of empirical research has been devoted to training convolutional rectifier networks. We conjecture that directing similar efforts into training convolutional arithmetic circuits, thereby fulfilling their expressive potential, may give rise to a deep learning architecture that is provably superior to convolutional rectifier networks yet has so far been overlooked by practitioners.

As discussed in the beginning of this section, depth efficiency alone does not unravel the mystery behind the expressive power of deep convolutional networks. For a complete understanding of the latter, one must consider the inductive bias, i.e. the properties of functions realized by polynomially-sized deep networks, and their suitability for real-world tasks. This is the purpose of the work in [5], where we study the ability of convolutional arithmetic circuits to model correlations among regions of their input. Correlations are formalized through the notion of separation rank, which for a given input partition, measures how far a function is from being separable. We show that a polynomially-sized deep network supports exponentially high separation ranks for certain input partitions, while being limited to polynomial separation ranks for others. The network's pooling geometry

effectively determines which input partitions are favored, thus serves as a means for controlling the inductive bias. Contiguous pooling windows as commonly employed in practice favor interleaved (entangled) partitions over coarse ones, orienting the inductive bias towards the statistics of natural images. Other pooling geometries lead to different preferences, and this allows tailoring convolutional networks for new types of data that depart from the usual domain of natural imagery. We validate this empirically with both convolutional arithmetic circuits and convolutional rectifier networks, showing that for image processing tasks of a local nature, such as characterization of shape continuity, standard contiguous pooling is optimal. On the other hand, for tasks such as symmetry detection, where modeling correlations between distinct input regions is important, scattered pooling geometries lead to better performance.

The prescription for tailoring a network to model correlations needed for a given task, is an exemplar of how our theory, developed to address fundamental questions regarding the expressiveness of convolutional networks, also brings forth new capabilities to their application in practice. We take this further in the next section, where I discuss two works leveraging our theory for designing new types of networks with novel capabilities and improved performance.

Practical Applications. Convolutional arithmetic circuits, born by our construction of a universal hypotheses space equipped with hierarchical tensor decompositions, are in fact closely related to probabilistic generative models. Namely, if the weights of each filter are constrained to lie on the simplex (non-negative and sum to one), the computation carried out by a convolutional arithmetic circuit produces the likelihood of the input under a universal high-dimensional generative model. As opposed to other generative methods recently considered in the literature (e.g. generative adversarial networks or variational models), our model admits tractable inference (computation of likelihood), and more importantly, tractable marginalization. This allows for previously infeasible capabilities such as classification under missing data where the missingness distribution at test time is unknown. We demonstrate this on image recognition benchmarks in [6].

The equivalence we established between convolutional networks and tensor decompositions applies in particular to *dilated convolutional networks* – a newly introduced variant that provides state of the art performance in audio and text processing tasks. With dilated convolutional networks, the choice of dilations throughout a network corresponds to determination of the mode tree underlying the respective decomposition. We utilize this in [7], and introduce the notion of a *mixed tensor decomposition*, blending together multiple mode trees. Mixed tensor decompositions correspond to *mixed dilated convolutional networks*, formed by interconnecting hidden layers of networks with different dilations. We show that mixing decompositions allows representation of tensors much more efficiently than what would have been possible without the mixture, and by this prove that interconnecting dilated convolutional networks brings forth a boost to their expressiveness. Empirical evaluations demonstrate that this translates to significant gains in accuracy.

Future Work. There are various promising avenues for future research. One direction we are investigating is an extension of the equivalence with tensor decompositions beyond convolutional networks. Specifically, we have recently learned that a decomposition named *Tensor Train (TT)* can be viewed as equivalent to recurrent neural networks, opening the door to analyzing the expressive properties of the latter, as well as comparing them to convolutional networks. An additional path we are exploring is the relation between our theory and that of tensor networks in quantum mechanics. The latter was developed before hierarchical tensor decompositions were introduced, leading us to believe that it may be possible to suggest more efficient algorithms than those used today, and perhaps even shed some light on physical principles. Finally, in the longer term, I am interested in leveraging the equivalence between convolutional networks and tensor decompositions for addressing the fundamental theoretical questions beyond expressiveness – optimization and generalization.

References

[1] N. Cohen and A. Shashua, *SimNets: A Generalization of Convolutional Networks*, Advances in Neural Information Processing Systems (NIPS), Workshop on Deep Learning and Representation Learning, 2014.
[2] N. Cohen, O. Sharir and A. Shashua, *Deep SimNets*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[3] N. Cohen, O. Sharir and A. Shashua, *On the Expressive Power of Deep Learning: A Tensor Analysis*, Conference On Learning Theory (COLT), 2016.
[4] N. Cohen and A. Shashua, *Convolutional Rectifier Networks as Generalized Tensor Decompositions*, International Conference on Machine Learning (ICML), 2016.
[5] N. Cohen and A. Shashua, *Inductive Bias of Deep Convolutional Networks through Pooling Geometry*, International Conference on Learning Representations (ICLR), 2017.
[6] O. Sharir, R. Tamari, N. Cohen and A. Shashua, *Tensorial Mixture Models*, arXiv preprint, 2017.
[7] N. Cohen, R. Tamari and A. Shashua, *Boosting Dilated Convolutional Networks with Mixed Tensor Decompositions*, International Conference on Learning Representations (ICLR), 2018.

Robust one-bit compressed sensing with non-Gaussian matrices

Sjoerd Dirksen
(joint work with Hans Christian Jung, Shahar Mendelson, Holger Rauhut)

The theory of compressed sensing predicts that one can reconstruct signals from a small number of linear measurements using efficient algorithms, by exploiting the empirical fact that many real-world signals possess a sparse representation. In the traditional compressed sensing literature, it is typically assumed that one can reconstruct a signal based on its analog linear measurements. In a realistic sensing scenario, measurements need to be quantized to a finite number of bits before they can be transmitted, stored, and processed. Formally, this means that one needs to reconstruct a sparse signal x based on *non-linear* observations of the

form $y = Q(Ax)$, where $Q : \mathbb{R}^m \to \mathcal{A}^m$ is a quantizer and $\mathcal{A}$ denotes a finite quantization alphabet.

We consider the *one-bit compressed sensing* model, which was first studied by Boufounos and Baraniuk [3]. In this model one observes

$$y = \operatorname{sign}(Ax + \tau), \tag{1}$$

where $A \in \mathbb{R}^{m \times N}$, $m \ll N$, sign denotes the sign function applied entry-wise and $\tau \in \mathbb{R}^m$ is a vector consisting of thresholds. Especially interesting is the *memoryless* one-bit quantization model, in which every linear measurement is quantized independently of the other measurements. The memoryless one-bit quantizer is attractive from a practical point of view, as it can be implemented using an energy-efficient comparator to a fixed voltage level (if the thresholds τ_i are all equal to a fixed constant) combined with dithering (if τ is random). In the original work [3] all thresholds were taken equal to zero. In this scenario, the energy of the original signal is lost during quantization and one can only hope to recover its direction.

There is by now a rich theory available for one-bit compressed sensing with standard Gaussian matrices. For instance, it is known that with high probability one can accurately reconstruct the direction of any (approximately) sparse signal via a tractable convex program, even if a fraction of the bits is corrupted at the quantizer in an adversarial manner [8]. This result is valid if the number of one-bit measurements m scales in terms of the signal sparsity s as $m \geq Cs \log(n/s)$, which is the optimal scaling known from 'unquantized' compressed sensing. More recently, it has been shown to be possible to efficiently reconstruct the complete signal by using Gaussian thresholds, provided that one knows an a-priori bound on the energy of the signal [2, 7]. Although these results are very interesting from a mathematical perspective, their practical value is limited by the fact that Gaussian matrices cannot be realized in a real-world measurement setup. It is therefore of substantial interest to extend the known theory to non-Gaussian matrices. This is a non-trivial task, as there exist measurement matrices that perform optimally in unquantized compressed sensing, but may fail if one-bit quantization is used. For instance, as was pointed out in [1], if A is a Bernoulli matrix and $\tau = 0$, then there already exist 2-sparse vectors that cannot be reconstructed based on their one-bit measurements, regardless of how many measurements we take. Still, [1] established a positive recovery result for subgaussian measurement matrices which shows that one can reconstruct a sparse signal x up to accuracy (at most) $\|x\|_\infty^{1/4}$. Informally, this means that one can still hope to recover the signal if it is sparse, but not too sparse.

In joint work with H.C. Jung and H. Rauhut [4], we establish the first rigorous reconstruction guarantees for memoryless one-bit compressed sensing with a structured random matrix. We investigate a randomly subsampled Gaussian circulant matrix, a measurement model that is relevant for several applications such as SAR radar imaging, Fourier optical imaging and channel estimation (see e.g. [9] and the references therein). In contrast to [1], the main results of [4] impose a *small sparsity assumption*. Under this assumption, we establish guarantees for

the accurate recovery of the direction of any s-sparse or effectively sparse vector using a single hard thresholding step or a linear program, respectively, in the case that the threshold vector τ is zero. Our analysis relies on work of S. Foucart [6], who observed that recovery results for these two reconstruction methods can be obtained by showing that the matrix A satisfies an ℓ_1/ℓ_2-restricted isometry property. By taking τ to be an appropriately scaled Gaussian vector, one can fully recover effectively sparse signals via a second order cone program, provided that an upper bound on their energy is known.

The works [1, 4] give the impression that in a non-Gaussian context one needs additional restrictions on the signal in order to accurately recover from one-bit measurements. In very recent joint work with S. Mendelson [5], we show that this impression is misleading. We prove that if one chooses the random threshold vector τ appropriately, then one can accurately reconstruct signals from general low-complexity sets based on their subgaussian, or even heavy-tailed, one-bit measurements. In the special case of sparse signals we additionally prove recovery results for randomly subsampled circulant matrices generated by a subgaussian vector, without any restriction on the sparsity level. In addition, our recovery results strongly improve over [1, 4] in terms of robustness: recovery is stable in the presence of adversarial bit corruptions in the quantization process, as well as heavy-tailed noise on the analog measurements. In the case of subgaussian and randomly subsampled subgaussian circulant matrices, robust recovery can be achieved via a convex (and, for many signal sets, tractable) recovery program.

References

[1] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. One-bit compressed sensing with non-Gaussian measurements. *Linear Algebra Appl.*, 441:222–239, 2014.

[2] R. G. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters. Exponential decay of reconstruction error from binary measurements of sparse signals. *IEEE Transactions on Information Theory*, 63(6):3368–3385, 2017.

[3] P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pages 16–21. IEEE, 2008.

[4] S. Dirksen, H. C. Jung, and H. Rauhut. One-bit compressed sensing with Gaussian circulant matrices. *ArXiv:1710.03287*, 2017.

[5] S. Dirksen and S. Mendelson. Robust one-bit compressed sensing with non-Gaussian measurements. *in preparation*, 2018.

[6] S. Foucart. *Flavors of Compressive Sensing*, pages 61–104. Springer International Publishing, Cham, 2017.

[7] K. Knudson, R. Saab, and R. Ward. One-bit compressive sensing with norm estimation. *IEEE Trans. Inform. Theory*, 62(5):2748–2758, 2016.

[8] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. *IEEE Trans. Inform. Theory*, 59(1):482–494, 2013.

[9] J. Romberg. Compressive sensing by random convolution. *SIAM J. Imaging Sci.*, 2(4):1098–1128, 2009.

Transports on manifolds in data analysis

Ronen Talmon
(joint work with Or Yair, Mirela Ben-Chen)

Consider two subsets $\left\{\boldsymbol{x}_i^{(1)}(t)\right\}_{i=1}^{N_1}$ and $\left\{\boldsymbol{x}_i^{(2)}(t)\right\}_{i=1}^{N_2}$, where $\boldsymbol{x}_i^{(k)}(t) \in \mathbb{R}^D$, of N_1 and N_2 high-dimensional time series, respectively. Assume that each subset is obtained from the same acquisition system in a particular session, deployment, and set of environmental conditions. In our notation, the superscript denotes the index of the subset, the subscript i denotes the index of the time-series within each subset, and t represents the time axis of each time-series. Our exposition focuses only on two subsets, but generalization to any number of subsets is straightforward. Additionally, we consider here time-series, but our derivation does not take the temporal order into account, and therefore, extension to other types of data, e.g., images, is immediate, where t could be just an index of a sample.

Analyzing such data typically raises many challenges. For example, one notable problem is how to efficiently compare between high-dimensional point clouds, and particularly, time-series. When the data are real measured signals, sample comparisons become even more challenging, since such high-dimensional measured data usually contain high levels of noise.

In particular, in our setting, we face an additional challenge, since the data is given in two separate subsets. Comparing time-series from the same subset is a difficult task by itself, even more so is comparing time-series from two different subsets.

Our goal in this work is to find a new *joint* representation of the two subsets in an unsupervised manner. Broadly, we aim to devise a low-dimensional representation in a Euclidean space that facilitates efficient and meaningful comparisons. As in many unsupervised tasks, this general description of the goal is not well-defined. To make our objective more concrete, we associate each time-series $\boldsymbol{x}_i^{(1)}(t)$ with a label $y_i^{(1)}$ and $\boldsymbol{x}_i^{(2)}(t)$ with a label $y_i^{(2)}$, and define "meaningful" comparisons with respect to these labels. Namely, we design the new representation such that the Euclidean distance between the new representations of any two time-series with similar corresponding labels is small, independently of the time-series respective trial, and particularly, subset. *To construct such a representation, we propose an approach that computes covariance matrices as data features, and then employs parallel transport on the manifold of symmetric positive-definite matrices* [1, 2, 3].

Based on our new representation, we devise efficient and accurate solutions for transfer learning and domain adaptation, which are long-standing problems in data analysis. Specifically, given a subset $\left\{\boldsymbol{x}_i^{(1)}(t)\right\}_{i=1}^{N_1}$ with corresponding labels $\left\{y_i^{(1)}\right\}_{i=1}^{N_1}$, we train a classifier on the new derived representation of the subset. Then, when another unlabeled subset $\left\{\boldsymbol{x}_i^{(2)}(t)\right\}_{i=1}^{N_1}$ becomes available, we apply the trained classifier to the derived (joint) representation.

To put the problem setting and our proposed solution in context, we will use an illustrative example, taken from a recent competition (`http://www.bbci.de/competition/iv/`). Consider data from a brain computer interface (BCI) experiment of motor imagery comprising D Electroencephalography (EEG) recordings. In this experiment, several subjects were asked to repeatedly perform one out of four motor imagery tasks (to raise their right hand, left hand, or feet, or to move their tongue), Let $\left\{\boldsymbol{x}_i^{(1)}(t)\right\}_{i=1}^{N_1}$ be a subset of recordings acquired from a single subject, indexed (1), where the time-series $\boldsymbol{x}_i^{(1)}(t)$ consists of the signals, recorded simultaneously from the D EEG channels during the ith trial. Each time series $\boldsymbol{x}_i^{(1)}(t)$ is attached with a label $y_i^{(1)}$, denoting the imagery task performed at the the ith trial. Common practice is to train a classifier based on $\left\{\boldsymbol{x}_i^{(1)}(t)\right\}_{i=1}^{N_1}$ and $y_i^{(1)}$, so that the imagery task could be identified from new EEG recordings. This capability could then be the basis for devising brain computer interfaces, for example, to control prosthetics.

Suppose that a new subset $\left\{\boldsymbol{x}_i^{(2)}(t)\right\}_{i=1}^{N_1}$ of recordings acquired from another subject, indexed (2), becomes available. Applying the classifier, trained based on data from subject (1), to the new subset of recordings from subject (2) typically yields poor results, as we demonstrate in our study. Indeed, to the best of our knowledge, all methods addressing this particular dataset, e.g., [4], as well as other related problems, exclusively analyze data from each individual subject separately. By constructing a joint representation for both $\left\{\boldsymbol{x}_i^{(1)}(t)\right\}_{i=1}^{N_1}$ and $\left\{\boldsymbol{x}_i^{(2)}(t)\right\}_{i=1}^{N_1}$, which is oblivious to the specific subject, we are able to build a classifier that is trained on data from one subject and applied to data from another subject without any calibration, i.e., without any labeled data from the new (test) subject.

References

[1] X. Pennec, P. Fillard, and N. Ayache, *A riemannian framework for tensor computing*, *International Journal of computer vision*, vol. 66, no. 1, pp. 41–66, 2006.

[2] R. Bhatia, *Positive definite matrices*. Princeton university press, 2009.

[3] S. Sra and R. Hosseini, *Conic geometric optimization on the manifold of positive definite matrices*, *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 713–739, 2015.

[4] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, *Classification of covariance matrices using a riemannian-based kernel for bci applications*, *Neurocomputing*, vol. 112, pp. 172–178, 2013.

Deep networks: engineered, trained, or randomized?

Rémi Gribonval

Many of the data analysis and processing pipelines that have been carefully engineered by generations of mathematicians and practitioners can in fact be implemented as deep networks. Allowing the parameters of these networks to be automatically trained allows to revisit certain constructions.

The talk first describes an empirical approach to approximate a given matrix by a fast linear transform through numerical optimization [1]. The main idea is to write fast linear transforms as products of few sparse factors, and to iteratively optimize over the factors. This corresponds to training a linear multilayer neural network with sparse connections. Algorithms exploiting iterative hard-thresholding projections have been shown to perform well in practice. Yet, developing a solid understanding of their conditions of success remains an open mathematical question.

In a second part, the talk outlines the main features of a recent framework for large-scale learning called compressive statistical learning [2]. Inspired by compressive sensing, the framework allows drastic volume and dimension reduction to learn from large/distributed/streamed data collections. Its principle is to compute a low-dimensional (nonlinear) sketch (a vector of random empirical generalized moments), in essentially one pass on the training collection.

For certain learning problems such as clustering [3], small sketches have been shown to capture the information relevant to the considered learning task, and empirical learning algorithms have been proposed to learn from such sketches. As a proof of concept, more than a thousands hours of speech recordings can be distilled to a sketch of only a few kilo-bytes, while capturing enough information estimate a Gaussian Mixture Model for speaker verification [4].

The framework, which is endowed with statistical guarantees in terms of learning error, is illustrated on sketched clustering, and sketched PCA, using empirical algorithms inspired by sparse recovery algorithms used in compressive sensing. The promises of the framework in terms of privacy-aware learning are discussed, as well as its connections with information preservation along pooling layers of certain convolutional neural networks with random weights.

To conclude the talk, we describe ongoing work [5] providing definitions and some characterizations of the approximation spaces [6] of deep networks and their relations with classical function spaces. Of particular interest is the role of the so-called activation function, and that of the depth of the considered networks.

References

[1] L. Le Magoarou, R. Gribonval, *Flexible Multi-layer Sparse Approximations of Matrices and Applications*, IEEE Journal of Selected Topics in Signal Processing **10 (4)** (2016).

[2] R. Gribonval, G. Blanchard, N. Keriven, Y. Traonmilin, *Compressive Statistical Learning with Random Feature Moments*, preprint, 2017.

[3] N. Keriven, N. Tremblay, Y. Traonmilin, R. Gribonval, *Compressive K-means*, International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017).

[4] N. Keriven, A. Bourrier, R. Gribonval, P. Perez, *Sketching for Large-Scale Learning of Mixture Models*, Information and Inference, 2017
[5] R. Gribonval, G. Kutyniok, M. Nielsen, F. Voigländer, *Approximation spaces of deep neural networks*, working draft.
[6] R. A. DeVore, G. G. Lorentz, *Constructive approximation*, 1993, Springer-Verlag.

Approximation Properties of Deep ReLU Networks

Felix Voigtlaender
(joint work with Philipp Petersen)

In the area of machine learning, deep learning methods have dramatically improved the state-of-the-art in many classification problems like visual object recognition [5]. The general goal of machine learning is to find a good approximation f^* to an *unknown* ground-truth (classifier) function f that can only be observed through known samples $(x_i, f(x_i))_{i=1,\dots,N}$. In the case of deep learning this is achieved by insisting that f^* is implemented by a **neural network** $\Phi = \Phi_a$ which is parametrized by its weights $a \in \mathbb{R}^K$. To determine these weights, one applies a form of stochastic gradient descent in order to minimize a loss function L that is defined in terms of the samples $(x_i, f(x_i))_{i=1,\dots,N}$. For details we refer to [5, 9].

Despite their incredible performance in applications, a theoretical explanation for this success of deep learning methods is still missing. In this abstract, we present recent results concerning the expressive power of neural networks. In particular, our results partially explain why *deep networks tend to perform better than shallow ones*, as is observed in practice [5].

We emphasize that we are only interested in the *existence* of a network Φ_ε^f approximating a given ground-truth classifier function f up to error ε. We do *not* address the practically important question of how one can *find* such a network, much less if one is only given samples of f.

1. Classical results

A neural network $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ computes its output by alternatingly applying affine-linear maps and a non-linear **activation function** $\varrho : \mathbb{R} \to \mathbb{R}$; thus

$$\Phi(x) = T_L(\varrho(T_{L-1}(\dots \varrho(T_1(x)) \dots))) \quad \text{for} \quad x \in \mathbb{R}^{N_0},$$

where $L \in \mathbb{N}$ denotes the **depth** of the network, and where each $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ is affine-linear, say $T_\ell = A_\ell \bullet + b_\ell$. Note that ϱ *is applied componentwise*. If we want to emphasize the choice of ϱ, we say that Φ is a ϱ-network.

Observe that a 1-layer network is simply an affine-linear map, while a 2-layer network is a linear combination of ridge functions, i.e., $\Phi(x) = \sum_{i=1}^K \varrho(\langle x, a_i\rangle + b_i)$.

The **number of neurons** and the **number of weights** of Φ are, respectively,

$$N(\Phi) = \sum_{\ell=0}^{L} N_\ell \quad \text{and} \quad W(\Phi) = \sum_{\ell=1}^{L} (\|A_\ell\|_{\ell^0} + \|b_\ell\|_{\ell^0}).$$

For later use, we recall the notion of sigmoidal activation functions: A (measurable, locally bounded) function $\varrho : \mathbb{R} \to \mathbb{R}$ is **sigmoidal of order** $k \in \mathbb{N}_0$ if $\lim_{x\to\infty} \frac{\varrho(x)}{x^k} = 1$ and $\lim_{x\to-\infty} \frac{\varrho(x)}{x^k} = 0$. The so-called **ReLU (rectified linear unit) activation function** $\varrho_0(x) := x_+$ is sigmoidal of order 1.

The expressiveness of neural networks is a well-studied subject [1, 2, 3, 4, 6, 7]. In particular, the following **universal approximation theorem** seems to settle—at first glance—the question of the expressiveness of neural networks:

Theorem (cf. [6, Theorem 1]) If ϱ is continuous but not a polynomial and $K \subset \mathbb{R}^d$ is compact, then the family of 2-layer neural networks is dense in $C(K)$.

Note, however, that the theorem does *not* yield any bounds on the complexity of the networks Φ_ε^f that are used to approximate f up to error ε. Deriving such bounds for ReLU networks under suitable assumptions on f is our main goal. For certain other types of activation functions, such bounds are classical; for example:

1) In [7] it is shown that if ϱ is sigmoidal of order $k \geq 2$ and if $f \in C^s([0,1]^d)$, then $\|f-\Phi_n\|_{L^\infty} \lesssim n^{-s/d}$ for a ϱ-network with $N(\Phi_n) = n$ and $L(\Phi_n) = L(d, s, k)$.

2) In [2] it is shown for order zero sigmoidal activation functions ϱ that if one assumes that the Fourier transform of $f : \mathbb{R}^d \to \mathbb{C}$ has a finite first moment, then $\|f - \Phi_n\|^2_{L^2(\mu)} \lesssim n^{-1}$ for a 2-layer ϱ-network Φ_n with $N(\Phi_n) = n$. Here, μ is a fixed (but arbitrary) probability measure on $\mathbb{R}^d$ with compact support.

2. Results for ReLU networks

Most classical results about the approximation properties of neural networks do *not* apply to ReLU networks. Since in practice the ReLU is the most widely used activation function [5], these networks received much attention in recent years.

Yarotsky [13] showed for Sobolev functions $f \in W^{k,\infty}([0,1]^d)$ that there are ReLU networks Φ_ε^f of size $N(\Phi_\varepsilon^f) \lesssim W(\Phi_\varepsilon^f) \lesssim \varepsilon^{-d/k}$ and depth $L(\Phi_\varepsilon^f) \lesssim \ln(1/\varepsilon)$ satisfying $\|f - \Phi_\varepsilon^f\|_{L^\infty} \lesssim \varepsilon$. Thus, the depth of the networks Φ_ε^f tends to infinity if the approximation accuracy gets better. As far as we know, *if one insists on L^∞ approximation with the same "network size to approximation error" relation as above, it is unknown whether one can avoid this growth of the depth.* But for L^p approximation with $p < \infty$ the situation is different:

Theorem ([8, Theorem A.9]) There is some $c > 0$ such that for any $\varepsilon, p, \beta \in (0, \infty)$ and $f \in C^\beta([0,1]^d)$, we have $\|f - \Phi_\varepsilon^f\|_{L^p([0,1]^d)} \leq \varepsilon$ for a suitable ReLU network Φ_ε^f satisfying $N(\Phi_\varepsilon^f) \lesssim W(\Phi_\varepsilon^f) \lesssim \varepsilon^{-d/\beta}$ and $L(\Phi_\varepsilon^f) \leq c \cdot (1 + \beta/d) \cdot \log_2(2 + \beta)$.

The ground-truth classifier function f, however, usually has a discrete range; e.g. in a digit classification problem, we could have $f : \mathbb{R}^{N\times N} \to \{-1, 0, \dots, 9\}$, where -1 stands for "not a number". Since such a function cannot be smooth, we consider a different "toy-model" for classifiers f, namely $f = \sum_{i=1}^M a_i \mathbb{1}_{K_i}$, where the sets $K_i \subset \mathbb{R}^d$ are assumed to have a smooth boundary, say $\partial K_i \in C^\beta$. By using that locally—after a change of variables—the functions $\mathbb{1}_{K_i}$ are similar to

jumps along straight lines, and that ReLU networks can approximate such jumps well, e.g. by $\varepsilon^{-1} \cdot (\varrho_0(x_i) - \varrho_0(x_i - \varepsilon))$, we proved the following result:

Theorem ([8, Theorem 3.5]) There is $c > 0$ such that for any $f = \sum_{i=1}^{M} a_i \mathbb{1}_{K_i}$ with $\partial K_i \in C^\beta$ and any $\varepsilon > 0$ there is a ReLU network Φ_ε^f with $\|f - \Phi_\varepsilon^f\|_{L^2([0,1]^d)} \leq \varepsilon$ and $N(\Phi_\varepsilon^f) \lesssim W(\Phi_\varepsilon^f) \lesssim \varepsilon^{-2(d-1)/\beta}$ as well as $L(\Phi_\varepsilon^f) \leq c \cdot \log_2(2+\beta) \cdot (1 + \beta/d)$.

Using entropy arguments, we showed that the bound $W(\Phi_\varepsilon^f) \lesssim \varepsilon^{-2(d-1)/\beta}$ is *optimal*, assuming that each of the weights of the network can be encoded with $\lesssim \varepsilon^{-t}$ bits for a fixed $t > 0$. We refer to [8, Section 4.1] for the details.

Telgarsky [12, 11] observed for ReLU networks $\Phi : \mathbb{R}^d \to \mathbb{R}$ that the one-dimensional restrictions $\mathbb{R} \to \mathbb{R}, t \mapsto \Phi(ta + b)$ (with $a, b \in \mathbb{R}^d$) are piecewise affine-linear with at most $N(\Phi)^{L(\Phi)}$ pieces, and used this to show that there are deep neural networks of size n that can only be approximated by shallow networks whose size is exponential in n. Utilizing this observation, we were able to prove the following lower bound for the approximation of non-linear smooth functions:

Theorem ([8, Theorem 4.5], see [10, Theorem 4] for the case $p = 2$)
Let $\varnothing \neq \Omega \subset \mathbb{R}^d$ be open and connected and let $f \in C^3(\Omega)$ not affine-linear. Then there is a constant $C_f > 0$ such that for every $p \in [1, \infty]$,

$$\|f - \Phi\|_{L^p(\Omega)} \geq C_f \cdot \max\left\{(N(\Phi) - 1)^{-2L(\Phi)}, \quad (W(\Phi) + d)^{-2L(\Phi)}\right\}.$$

Overall, our results show that *smoother functions allow for better approximation rates by ReLU networks; but to achieve these rates, deep networks are needed!*

References

[1] M. Anthony and P.L. Bartlett, *Neural network learning: Theoretical foundations*, Cambridge University Press, Cambridge, 1999.

[2] A.R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Transactions on Information Theory **39(3)**, (1993), 930–945.

[3] G. Cybenko, *Approximations by superpositions of sigmoidal functions*, Mathematics of Control, Signals, and Systems **2(4)** (1989), 303–314.

[4] K. Hornik, *Approximation capabilities of multilayer feedforward networks*, Neural Networks **4(2)** (1991), 251–257.

[5] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature **521**, (2015), 436–444.

[6] M. Leshno, V.Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, Neural Networks **6**, (1993), 861–867.

[7] H. N. Mhaskar, *Approximation properties of a multilayered feedforward artificial neural network*, Advances in Computational Mathematics **1(1)**, (1993), 61–80.

[8] P. Petersen and F. Voigtlaender, *Optimal approximation of piecewise smooth functions using deep ReLU neural networks*, arXiv preprints, `arxiv.org/abs/1709.05289`, 2017.

[9] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, *Learning representations by back-propagating errors*, Nature **323**, (1986), 533–536.

[10] I. Safran and O. Shamir, *Depth-width tradeoffs in approximating natural functions with neural networks*, Proceedings of Machine Learning Research, **70**, (2017), 2979–2987.

[11] M. Telgarsky, *Representation benefits of deep feedforward networks*, arXiv preprints, `https://arxiv.org/abs/1509.08101`, 2015.
[12] M. Telgarsky, *Benefits of depth in neural networks*, Conference on Learning Theory (COLT), (2016), 1517–1539.
[13] D. Yarotsky, *Error bounds for approximations with deep ReLU networks*, Neural Networks **94** (2017), 103–114.

Time evolving data, diffusion geometry, and randomized matrix decomposition

Nicholas F. Marshall
(joint work with Matthew J. Hirn [5])

We describe how the geometry of time evolving data can be efficiently summarized using diffusion operators and randomized matrix decomposition. Suppose than an $n \times d \times m$ tensor corresponding to n points in $\mathbb{R}^d$ measured over m times is given. For each $n \times d$ temporal slice X_i of the tensor we construct a diffusion operator P_i, following the diffusion maps framework [1], and study the product operator

$$P^{(m)} := P_m P_{m-1} \cdots P_1.$$

We prove that this product operator approximates heat flow in a precise sense when a manifold with a time dependent metric is assumed to underlie the data. Furthermore, we generalize the notion of diffusion distance and diffusion maps to this time evolving setting. We observe that the singular value decomposition of the product operator $P^{(m)}$ can be efficiently computed by implementing each P_i as a sparse matrix, applying each matrix successively to a collection of random vectors, and then using the algorithm of Martinsson, Rokhlin, and Tygert [6]. This decomposition can in turn be used to compute a generalized diffusion map, which we call a time coupled diffusion map, that summarizes the geometry of the data tensor. We remark that other recent works in the diffusion geometry literature also consider embeddings defined via products of diffusion kernels, see for example Lederman and Talmon [3], or Lindenbaum, Yeredor, Salhov, and Averbuch [4].

Main result. Our main result establishes a connection between the product operator $P^{(m)}$ and heat flow on an assumed underlying manifold with a time dependent metric. The existence and uniqueness of the heat kernel H_0^t on a manifold with a time dependent Riemannian metric was established by Guenther [2]. In order to prove a convergence result, we introduce a dependence on a bandwidth parameter ε for our product operator and write $P^{(m)} = P_\varepsilon^{(m)}$. Recall that n is the number of spatial samples of the manifold $\mathcal{M}$, and that m is the number of temporal measurements. We assume that the underlying time interval $[0, T]$ is divided into m intervals $[\tau_{i-1}, \tau_i)$ each of length ε where $\tau_0 = 0$ and $\tau_m = T$. For simplicity, we assume that our m measurements are taken at $(\tau_1, \ldots, \tau_m)$. Our main result is that in the limit of large data, both spatially and temporally, the product operator $P_\varepsilon^{(\lceil t/\varepsilon \rceil)}$ converges to the heat kernel:

$$P_\varepsilon^{(\lceil t/\varepsilon \rceil)} \to H_0^t \text{ as } n \to \infty \text{ and } \varepsilon \to 0.$$

More precisely:

Theorem. *Suppose the isometric embedding $\mathcal{M}_\tau \subset \mathbb{R}^d$ of a time dependent manifold $(\mathcal{M}, g(\tau))$ is measured at a common set $X = \{x_j\}_{j=1}^n \subset \mathcal{M}$ of n points at ε spaced units of time over a time interval $[0,T]$, so that, in particular, we have time samples $(\tau_i)_{i=1}^m \subset [0,T]$ with $\tau_i = i \cdot \varepsilon$ and $m = T/\varepsilon$.*

Then, for any sufficiently smooth function $f : \mathcal{M} \to \mathbb{R}$ and $t \leq T$, the heat kernel H_0^t can be approximated by the operator $P_\varepsilon^{(\lceil t/\varepsilon \rceil)}$:

$$P_\varepsilon^{(\lceil t/\varepsilon \rceil)} f(x_j) = H_0^t f(x_j) + \mathcal{O}\left(\frac{1}{n^{1/2}\varepsilon^{d/4+1/2}}, \varepsilon\right), \quad x_j \in X.$$

Time coupled diffusion distance. Let δ_j denote a Dirac distribution centered at x_j. We compare the points x_j and x_k by comparing the posterior distributions of $\delta_j^\top$ and $\delta_k^\top$ under the Markov operator $P^{(m)}$. More specifically, following [1] we define a diffusion based distance as the L^2 distance between these posterior distributions weighted by the reciprocal of the stationary distribution of the Markov chain. That is, we define the distance $D^{(m)}$ by

$$D^{(m)}(x_j, x_k) = \|\delta_j^\top P^{(m)} - \delta_k^\top P^{(m)}\|_{L^2(1/\pi_{(m)})},$$

where $\pi_{(m)}$ is the stationary distribution of $P^{(m)}$, i.e., $\pi_{(m)}^\top = \pi_{(m)}^\top P^{(m)}$, and $\|\cdot\|_{L^2(1/\pi_{(m)})}$ is the weighted L^2 norm:

$$\|f\|_{L^2(1/\pi_{(m)})} := \sqrt{\sum_{j=1}^n f(x_j)^2 \frac{1}{\pi_{(m)}(x_j)}}.$$

Time coupled diffusion map. The product operator $P^{(m)}$ is not, in general, similar to a symmetric matrix (as in the standard diffusion maps framework [1]) so our definition of a diffusion map necessarily differs. First, we define the operator $A^{(m)}$ by

$$A^{(m)} = \Pi_{(m)}^{1/2} P^{(m)} \Pi_{(m)}^{-1/2},$$

where $\Pi_{(m)}$ denotes the matrix with the stationary distribution $\pi_{(m)}$ of $P^{(m)}$ along the diagonal and zeros elsewhere. Next, we compute the singular value decomposition (SVD) of $A^{(m)}$:

$$A^{(m)} = U_{(m)} \Sigma_{(m)} V_{(m)}^\top,$$

where $U_{(m)}$ is an orthogonal matrix of left singular vectors, $\Sigma_{(m)}$ is a diagonal matrix of corresponding singular values, and $V_{(m)}$ is an orthogonal matrix of right singular vectors. Define

$$\Psi^{(m)} := \Pi_{(m)}^{-1/2} U_{(m)} \Sigma_{(m)}.$$

Then it is easy to check (see [5]) that the embedding

$$x_j \mapsto \delta_j^\top \Psi^{(m)}$$

of the data X into Euclidean space preserves the time coupled diffusion distance. That is to say,

$$D^{(m)}(x_j, x_k) = \|\delta_j^\top \Psi^{(m)} - \delta_k^\top \Psi^{(m)}\|_{L^2}.$$

We refer to the embedding $x_j \mapsto \delta_j^\top \Psi^{(m)}$ as the time coupled diffusion map.

References

[1] R. R. Coifman, S. Lafon, Diffusion maps, Applied and Computational Harmonic Analysis 21 (1) (2006) 5 -30.

[2] C. M. Guenther, The fundamental solution on manifolds with time dependent metrics, The Journal of Geometric Analysis 12 (3) (2002) 425–436.

[3] R. R. Lederman, R. Talmon, Common manifold learning using alternating-diffusion, Tech. rep., Yale (2014).

[4] O. Lindenbaum, A. Yeredor, M. Salhov, A. Averbuch, Multiview diffusion maps, arXiv:1508.05550 (2015).

[5] N. F. Marshall and M. J. Hirn. Time coupled diffusion maps. Applied and Computational Harmonic Analysis (2017).

[6] P. G. Martinsson, V. Rokhlin, M. Tygert. A randomized algorithm for the decomposition of matrices. Applied and Computational Harmonic Analysis (2011).

Morphing of Manifold-Valued Images

Gabriele Steidl

(joint work with Sebastian Neumayer, Johannes Persch)

Smooth image transition, also known as image morphing, is a frequently addressed task in image processing and computer vision, and there are various approaches to tackle the problem. For example, in feature based morphing only specific features are mapped to each other and the whole deformation is then calculated by interpolation. This paper is related to a special kind of image morphing, the so-called metamorphosis introduced by Miller, Trouvé and Younes [4, 5]. The metamorphosis model can be considered as an extension of the flow of diffeomorphism model and its large deformation diffeomorphic metric mapping framework in which *each* image pixel is transported along a trajectory determined by a diffeomorphism path. As an extension the metamorphosis model allows the variation of image intensities along trajectories of the pixels.

This paper builds up on a time discrete geodesic paths model by Berkels, Effland and Rumpf [1], but considers images in $L^2(\Omega, \mathcal{H})$, where $\Omega \subset \mathbb{R}^n$, $n \geq 2$, is an open, bounded connected domain with Lipschitz boundary and $\mathcal{H}$ a finite dimensional Hadamard manifold. Hadamard manifolds are simply connected, complete Riemannian manifolds with non-positive sectional curvature. Typical examples are hyperbolic spaces and symmetric positive definite matrices with the affine invariant metric. As an important fact we will use that the distance in Hadamard spaces is jointly convex which will imply weak lower semicontinuity of certain functionals involving the distance function.

We aim in finding a minimizing sequence $\mathbf{I} = (I_1, \dots, I_{K-1}) \in \left(L^2(\Omega, \mathbb{R})\right)^{K-1}$ of the *discrete path energy*

$$\mathcal{J}(\mathbf{I}) := \sum_{k=1}^{K} \inf_{\varphi_k \in \mathcal{A}_\epsilon} \int_\Omega W(D\varphi_k(x)) + \gamma |D^m \varphi_k(x)|^2 \, dx + \frac{1}{\delta} \int_\Omega d_2(I_k \circ \varphi_k, I_{k-1})^2 \, dx,$$
$$\text{subject to} \quad I_0 = T, \; I_K = R, \tag{1}$$

where $\delta, \gamma > 0$, d_2 denotes the distance in $L^2(\Omega, \mathcal{H})$,

$$\mathcal{A}_\epsilon := \{\varphi \in \left(W^{m,2}(\Omega)\right)^n : \det(D\varphi) \geq \epsilon, \; \varphi(x) = x \text{ for } x \in \partial\Omega\}, \quad m > 1 + \frac{n}{2}$$

is an admissible set of deformations and the function W has to satisfy certain properties. A particular choice of W is given by the linearized elastic potential. An illustration is given in Fig. 1. We prove that a minimizer of (1) exists.

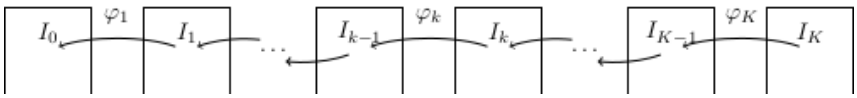

FIGURE 1. Illustration of the time discrete morphing path.

Dealing with digital images we have to introduce a space discrete model. We establish a finite difference model on a staggered grid together with a multiscale strategy. We have used this discretization already for gray-value images in [3]. For finding a minimizer, we also propose an alternating algorithm fixing either the deformation or the image sequence:

i) For a fixed image sequence, we have to solve certain registration problems for *manifold-valued images* in parallel to get a sequence $(\varphi_1, \dots, \varphi_K)$ of deformations. Necessary interpolations were performed via Karcher means computation.
ii) For a fixed deformation sequence, we need to find a minimizing image sequence $(I_1, \dots, I_{K-1})$ of

$$\sum_{k=1}^{K} \mathrm{d}_2^2(I_k \circ \varphi_k, I_{k-1}) \quad \text{subject to} \quad I_0 = T, I_K = R$$

where d_2 denotes the distance in $L^2(\Omega, \mathcal{H})$.

Fig. 2 shows a path obtained by our model for images with 3×3 positive definite matrices as entries. For more information we refer to [2].

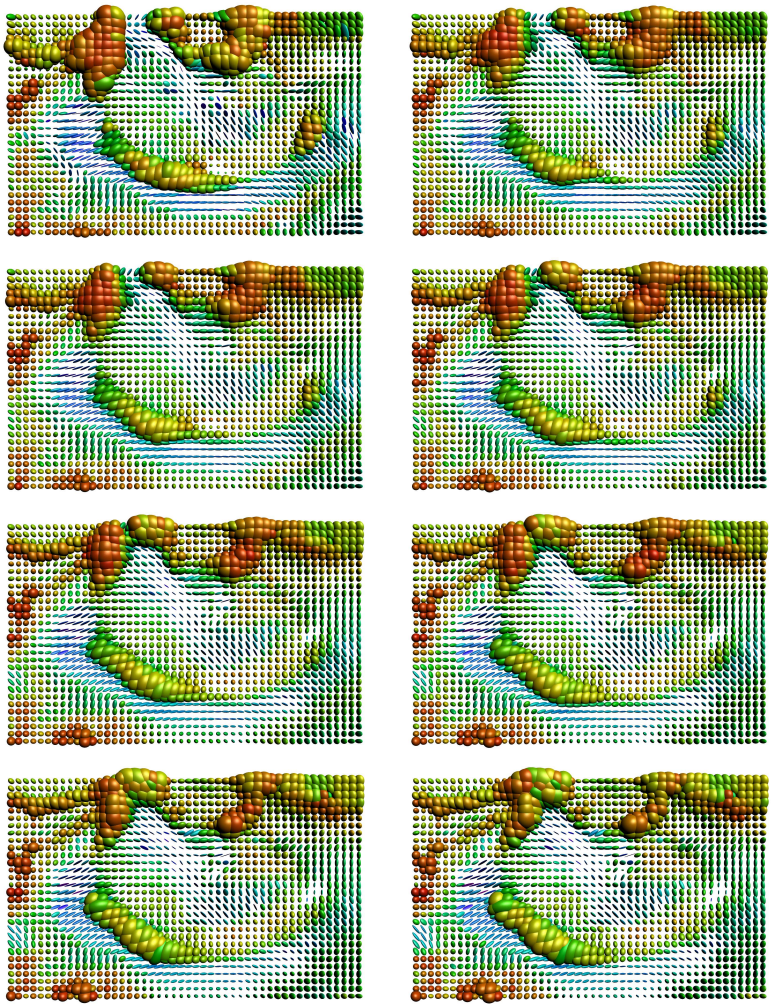

FIGURE 2. Morphing path between a part of the YZ-slices 49 and 51 of the Camino dataset with SPD(3) matrices.

REFERENCES

[1] B. Berkels, A. Effland, and M. Rumpf. Time discrete geodesic paths in the space of images. *SIAM Journal on Imaging Sciences* **8(3)** (2015), 1457–1488.

[2] J. Persch, F. Pierre, and G. Steidl. Exemplar-based face colorization using image morphing. *Journal of Imaging* **3(4)** (2017) :ArtNum 48.

[3] J. Persch, F. Pierre, and G. Steidl. Morphing of manifold-valued images inspired by discrete geodesics in image spaces. *SIAM Journal on Imaging Sciences*, submitted.
[4] A. Trouvé and L. Younes. Local geometry of deformable templates. *SIAM Journal of Mathematical Analysis* **37(2)** (2005), 17–59.
[5] A. Trouvé and L. Younes. Metamorphoses through Lie group action. *Foundations in Computational Mathematics* **5(2)** (2005), 173–198.

Solving linear Kolmogorov Equations by Means of Deep Learning

Philipp Grohs

1. The Mathematical Learning Problem

According to [1], the mathematical learning problem can be cast into the following form.

Definition 1 (The Mathematical Learning Problem)**.** *Let $K \subseteq \mathbb{R}^d$, let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space and let $X\colon \Omega \to K$ and $Y\colon \Omega \to \mathbb{R}^n$ be random vectors. For a Borel measurable function $F\colon K \to \mathbb{R}^n$ define the least squares error of F with respect to (w.r.t.) X and Y by*

$$\mathcal{E}_{(X,Y)}(F) = \int_\Omega \|F(X) - Y\|_{\mathbb{R}^n}^2 \, d\mathbb{P} = \mathbb{E}\big[\|F(X) - Y\|_{\mathbb{R}^n}^2\big] \in [0, \infty]. \tag{1}$$

The Mathematical Learning Problem asks for a function F which minimizes $\mathcal{E}_{(X,Y)}(F)$.

This definition can be interpreted as the problem of finding the best functional relation between two random vectors X, Y where X may take the role of a data point and Y that of a label. Since the minimization of $\mathcal{E}_{(X,Y)}$ amounts to a quadratic minimization problem, the solution to the Mathematical Learning Problem of Definition 1 can be easily seen to be the conditional expectation $\hat{F}(x) = \mathbb{E}(Y|X = x)$.

In practice, one does not know the distributions of (X, Y) but one only has access to i.i.d. samples $(x_i, y_i)_{i=1}^m \sim (X, Y)$ from which one needs to estimate $\hat{F}$. A popular method to achieve this goal is Empirical Risk Minimization (ERM), which minimizes the empirical risk

$$\mathcal{E}_{(x_i,y_i)_{i=1}^m}(F) := \frac{1}{m}\sum_{i=1}^m (F(x_i) - y_i)^2 \tag{2}$$

over a hypothesis class $\mathcal{H} \subset C(\mathbb{R}^d, \mathbb{R}^n)$. The minimizer (which may be non-unique) is denoted $\hat{F}_{(x_i,y_i)_{i=1}^m,\mathcal{H}}$. Classical statistical learning theory, as for example presented in [1] provides an estimate on the error

$$\mathcal{E}_{(X,Y)}(\hat{F}_{(x_i,y_i)_{i=1}^m,\mathcal{H}}) - \mathcal{E}_{(X,Y)}(\hat{F}) = \Big\|\hat{F}_{(x_i,y_i)_{i=1}^m,\mathcal{H}} - \hat{F}\Big\|^2_{L^2(K,d\mathbb{P}_X)}.$$

These estimates involve bounds on the approximation error $\left\|\hat{F}_{\mathcal{H}} - \hat{F}\right\|^2_{L^2(K,d\mathbb{P}_X)}$, where $\hat{F}_{\mathcal{H}} \in \operatorname{argmin}_{F\in\mathcal{H}} \|F - \hat{F}\|^2_{L^2(K,d\mathbb{P}_X)}$ and the generalization error $\mathcal{E}_{(X,Y)}(\hat{F}_{(x_i,y_i)_{i=1}^m,\mathcal{H}}) - \mathcal{E}_{(X,Y)}(\hat{F}_{\mathcal{H}})$. The approximation error measures how well the hypothesis class $\mathcal{H}$ approximates the regression function $\hat{F}$ and its estimation implicitly requires the knowledge of regularity properties of (X, Y) which is in general not available for realistic learning problems.

2. Neural Network Hypothesis Classes

In recent years spectacular successes have been achieved by using deep (artificial feedforward) neural networks as hypothesis classes [2]. These can be defined as follows.

Definition 2. *Let $L, N_0, \dots, N_L \in \mathbb{N}$. A neural network (NN) Φ with L layers is a finite sequence of matrix-vector tuples $\Phi := ((A_1, b_1), (A_2, b_2), \dots, (A_L, b_L)) \in \times_{l=1}^{L} \left(\mathbb{R}^{N_l \times N_{l-1}} \times \mathbb{R}^{N_l}\right)$. We refer to the sequence* $\operatorname{arch}(\Phi) := (N_0, N_1, \dots, N_L)$ *as the architecture of Φ and denote its input and output dimension by $d_{in}(\Phi) := N_0$ and $d_{out}(\Phi) := N_L$, respectively.*

Suppose $\sigma \in C(\mathbb{R}, \mathbb{R})$, then we define the realization of Φ with activation function σ as the map $R_\sigma(\Phi) \in C(\mathbb{R}^{N_0}, \mathbb{R}^{N_L})$ with $R_\sigma(\Phi)(x) = x_L$, where x_L is given by the following scheme:

$$x_0 := x, \quad x_l := \sigma(A_l x_{l-1} + b_l), \text{ for } l \in \{1, \dots, L-1\}, \quad x_L := A_L x_{L-1} + b_L.$$

Here, σ is understood component wise, i.e., $\sigma(y) = (\sigma(y_1), \dots, \sigma(y_m))$. Finally, we define $\operatorname{size}(\Phi) := \sum_{j=1}^{L} (\|A_j\|_{\ell^0} + \|b_j\|_{\ell^0})$ *and set*

$$\mathcal{H}^{\sigma}_{(N_0,\dots,N_L)} := \{R_\sigma(\Phi) : \ \operatorname{arch}(\phi) = (N_0, \dots, N_L)\} \subset C(\mathbb{R}^{N_0}, \mathbb{R}^{N_L}).$$

Examples of activation functions include the rectified linear unit $\operatorname{ReLU}(t) := (t)_+$ or the sigmoidal function $\operatorname{sig}(t) = \tanh(t/2)$. The resulting ERM problem (2) becomes non-linear and non-convex and it can typically only be solved by stochastic first order optimization methods whose convergence properties are not yet understood.

3. Linear Kolmogorov Equations as Learning Problem

Consider linear Kolmogorov equations which are defined as follows for a function $u : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ and initial value $\varphi : \mathbb{R}^d \to \mathbb{R}$:

(3)
$$\frac{\partial u}{\partial t}(t,x) = \frac{1}{2}\operatorname{Trace}(\sigma(x)\sigma^T(x)\operatorname{Hess}_x u(t,x)) + \mu(x) \cdot \nabla_x u(t,x), \ (t,x) \in [0,T] \times \mathbb{R}^d,$$
$$u(0,x) = \varphi(x).$$

The problem of approximating the function $x \mapsto u(T, x)$, $T > 0$, given the initial condition $u(0, x) = \varphi(x)$ arises in a wide area of applications, for example nonlinear filtering or computational finance. Special cases include diffusion equations,

the Black-Scholes-Equation or the Heston model which is used day after day in the financial engineering industry. In these applications it is especially relevant to develop efficient numerical schemes for high-dimensional problems with $d \geq 100$. Due to the curse of dimensionality which states that the complexity of classical methods (Finite Element Methods, Finite Difference Methods, Sparse Tensor Product Methods, Spectral Methods, ...) scales exponentially in the dimension d, such methods are not applicable in this regime.

In [3] use the Feynman-Kac formula $u(T,x) = \mathbb{E}(\varphi(Z_x^T))$, where Z_x^t is the process defined as $Z_x^t = x = \int_0^t \mu(Z_x^s)ds + \int_0^t \sigma(Z_x^2)dW_s$ to observe that $u(T,x)|_{[a,b]^d}$ equals the solution to the learning problem in the precise sense of Definition 1 associated with $X = \mathcal{U}_{[a,b]^d}$ (the uniform distribution on $[a,b]^d$) and $Y = \varphi(Z_X^t)$. Using this reformulation we can simulate training data $(x_i, y_i)_{i=1}^m$ distributed according to (X,Y) and solve the resulting ERM problem with a NN hypothesis class $\mathcal{H}_{(N_0,\dots,N_L)}^{\mathrm{ReLU}}$ which results in a numerical approximation $\hat{F}_{(x_i,y_i)_{i=1}^m, \mathcal{H}_{(N_0,\dots,N_L)}^{\mathrm{ReLU}}}$ of $\hat{F} = u(T,\cdot)$. Numerical simulations carried out in [3] suggest that the resulting algorithm does not suffer from the curse of dimensionality.

In ongoing work we show that in many cases of interest both the size $\sum_{l=1}^L (N_l \times N_{l-1} + N_l)$ of the NN hypothesis class as well as the number m of required training samples scale only polynomially in the dimension d.

Theorem 1 (informal and simplified version [G-Jentzen-von Wurstemberger])**.** *Suppose that μ, Σ are affine functions (this includes diffusion equations or the Black-Scholes equation) and suppose that the initial condition φ can be very well approximated by ReLU NNs (this includes $\varphi(x) = \max\{\sum_{i=1}^d x_i - K_i, 0\}$ (basket option) or $\varphi(x) = \max\{x_1 - K_1, \dots, x_d - K_d, 0\}$ (max option)). Then there exists a polynomial p such that for every $\epsilon > 0$ there exist $L, N_1, \dots, N_L, m \in \mathbb{N}$ with*

(i) $\sum_{l=1}^L (N_l \times N_{l-1} + N_l) \leq |p(d)|\epsilon^{-2}$
(ii) $m \leq |p(d)|\epsilon^{-4}$
(iii) $\frac{1}{(b-a)^d} \left(\int_{[a,b]^d} |u(T,x) - \hat{F}_{(x_i,y_i)_{i=1}^m, \mathcal{H}_{(N_0,\dots,N_L)}^{\mathrm{ReLU}}}(x)|^2 \right)^{1/2} \leq \epsilon.$

In other words, the method does not suffer from the curse of dimensionality.

References

[1] F. Cucker, S. Smale. *On the Mathematical Foundations of Learning*, Bulletin of the AMS **39**/1 (2001), 1–49.

[2] I. Goodfellow, Y. Bengio, A. Courville. *Deep Learning*, MIT press (2016).

[3] C. Beck, S. Becker, P. Grohs, N. Jaafari, A. Jentzen. *Solving stochastic differential equations and Kolmogorov equations by means of deep learning.* Preprint.

A new paradigm for function approximation with deep networks

Hrushikesh N. Mhaskar

A central problem in machine learning is the following. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^M \subset \mathbb{X} \times \mathbb{R}$ for some non-empty set $\mathbb{X}$. Find a model $P : \mathbb{X} \to \mathbb{R}$ such that $P(x_i) \approx y_i$, $1 \le i \le M$.

The traditional machine learning paradigm as described in [2, 3] is to consider $\mathcal{D}$ as an i.i.d. sample from an unknown probability distribution μ. The goal is to estimate the **generalization error** given by $\int_{\mathbb{X}\times\mathbb{R}} |y - P(x)|^2 d\mu(x, y)$. Writing $f(x) = \mathbb{E}_\mu(y|x)$, and denoting by μ^* the marginal distribution of μ for x, the generalization error is the sum of the **variance**, defined by $\int_{\mathbb{X}\times\mathbb{R}} |y-f(x)|^2 d\mu(x, y)$ and the **bias**, defined by $\int_{\mathbb{X}} |f(x) - P(x)|^2 d\mu^*(x)$. In theoretical analysis, one considers a sequence of model classes $V_0 \subset V_1 \subset \cdots$. The minimum bias for $P \in V_n$ is obtained for some $P^* = \arg\min_{P\in V_n} \|f - P\|_{\mu^*,2}$, and is called the **approximation error** $\|f - P^*\|^2_{\mu^*,2}$. In the traditional paradigm, the actual construction of P^* is of no interest; only an estimate of this minimum bias is studied in order to get some insight on what space V_n to choose the model from. The actual model $P^\#$ is computed based only on $\mathcal{D}$ typically using some empirical risk minimization process that assumes f to belong, for example, to a reproducing kernel Hilbert space (**prior**). Since the approximation error decreases as $n \uparrow \infty$, while the complexity of the process of finding $P^\#$ increases with n, there is an built-in trade off between the two estimates.

In the analysis of function approximation by deep networks, this paradigm does not work. As pointed out in [10], the main reason for deep networks to have an advantage over shallow networks is their ability to utilize a compositional structure so as to mitigate the curse of dimensionality with the blessing of compositionality. For example, the number of parameters in approximating a function F of 4 variables up to accuracy ϵ is $\mathcal{O}(\epsilon^{-r/4})$, where r measures the smoothness of F. However, if F has a compositional structure $F(x_1, \cdots, x_4) = f(f_1(x_1, x_2), f_2(x_3, x_4))$, then a deep network with binary tree architecture of the same form, $P(x_1, \cdots, x_4) = P^*(P_1(x_1, x_2), P_2(x_3, x_4))$ can provide the same approximation with only $\mathcal{O}(\epsilon^{-r/2})$ parameters, since only functions of two variables are approximated at each stage. If the functions f, f_1, f_2 are Lipschitz, one can easily obtain a rate of approximation using triangle inequality (**good propagation of errors**) (see [10] for details). This requires an approximation of $f(f_1, f_2)$ by $P^*(P_1, P_2)$ as **bivariate functions**. The inputs to f are thus different from those to its approximation P^*, and it is not possible to define a measure with respect to which to take the L^2 norm so that the measure is commensurate with the compositionality structure.

We propose an alternative, equivalent way to look at the problem that avoids the trade off between approximation error and process complexity, and utilizes the knowledge from approximation theory that can lead simultaneously to a good approximation error as well as an explicit construction for the desired model. Our

viewpoint is to treat the question of machine learning as the problem of finding a **good** approximation to an **unknown** function f that captures the essential functional relationship in the data set, given the values $y_i = f(x_i)+\epsilon_i$, where ϵ_i are i.i.d. samples drawn from an unknown distribution with mean 0 and independent of x_i. Although this is an equivalent formulation of the original problem, it is more natural from the point of view of approximation theory to do pointwise estimates (alternately uniform estimates involving some weight functions) than those in L^2, and more importantly to look for constructive methods that yield a good (rather than best) approximation. In particular, the generalization error is now defined as the pointwise error $|f(x) - P^{\#}(x)|$.

The author has developed constructive methods to accomplish this goal in many contexts (e.g. [5, 8, 9, 6, 7]). It is clear that these methods cannot yield an overall better accuracy than the L^2 projection methods when the error is measured in the global L^2 norm. In most applications though, the target function f is smooth on its domain $\mathbb{X}$ except for a small set of "singularities". It is well known in approximation theory that the error in L^2 projections are very sensitive to these singularities. In contrast, our methods produce errors according to the local smoothness of the target function at each point, analogous to those given in classical wavelet analysis [4, Chapter 9], in spite of the fact that they are defined using global data, with no a priori assumptions made on the smoothness of the target function whether globally or locally.

In our talk, available at `https://www.mathc.rwth-aachen.de/owncloud/index.php/s/GataT6XimZCWTwl`, we illustrated the local approximation properties of our methods in the context of function approximation on the Euclidean (hyper-)sphere. It is noted that approximation by ReLU networks on the Euclidean space can be reduced to an equivalent problem on the sphere. We also gave pointwise and uniform error estimates for shallow and deep networks to explain the phenomenon that it is possible to drive the training error to zero and yet keep the test error under control [1, 12, 11].

References

[1] M. Belkin, S. Ma, and S. Mandal, *To understand deep learning we need to understand kernel learning*, arXiv preprint arXiv:1802.01396, 2018.

[2] F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bulletin of the American Mathematical Society, **39** (2002), 1–49.

[3] F. Cucker and D. X. Zhou, *Learning theory: an approximation theory viewpoint*, **24** (2007), Cambridge University Press.

[4] I. Daubechies, *Ten lectures on wavelets*, **61** (1990), SIAM.

[5] Q. T. Le Gia and H. N. Mhaskar, *Localized linear polynomial operators and quadrature formulas on the sphere*, SIAM Journal on Numerical Analysis, **47**(1)(2008), 440–466.

[6] H. N. Mhaskar, *When is approximation by Gaussian networks necessarily a linear process?* Neural Networks, **17**(7) (2004), 989–1001.

[7] H. N. Mhaskar, *Weighted quadrature formulas and approximation by zonal function networks on the sphere*, Journal of Complexity, **22**(3) (2006), 348–370.

[8] H. N. Mhaskar, *Eignets for function approximation on manifolds*, Applied and Computational Harmonic Analysis, **29**(1) (2010), 63–87.

[9] H. N. Mhaskar, *Function approximation with relu-like zonal function networks*, arXiv preprint arXiv:1709.08174, 2017.
[10] H. N. Mhaskar and T. Poggio, *Deep vs. shallow networks: An approximation theory perspective*, Analysis and Applications, **14**(6) (2016), 829–848.
[11] H. N. Mhaskar and T. Poggio, *An analysis of training and generalization errors in shallow and deep networks*, arXiv preprint arXiv:1802.06266, 2018.
[12] T. Poggio, K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, and H. N. Mhaskar, *Theory of deep learning iii: explaining the non-overfitting puzzle*, arXiv preprint arXiv:1801.00173, 2017.

On the gap between local recovery guarantees in structured compressed sensing and oracle estimates

Claire Boyer

This is an on-going work with Ben Adcock and Simone Brugiapaglia (Simon Fraser University, Burnaby). Compressed sensing theory provides guarantees to reconstruct sparse signals from a few linear of measurements. However, in order to circumvent combinatorial issue, the "good" theoretical sensing matrices are often random, typically

(i) Gaussian matrices with i.i.d. Gaussian entries,
(ii) matrices obtained by stacking rows drawn from a finite-dimensional isometry, for instance randomly selected Fourier atoms.

However in practice, the acquisition is very structured due to the physics of acquisition, and measurements can be performed by groups or blocks, standing for admissible sampling patterns.

Therefore, in this work, we consider a compressed sensing (CS) theory more compatible with real-life applications: we derive guarantees to ensure reconstruction of a structured sparse signal of interest while imposing structure in the acquisition. We actually extend the setting of [1].

Once this setting established, one can study oracle-type bound: one can show that if we know the support of the signal to reconstruct, to ensure robust recovery, the required number of measurements can read as follows

$$m \geq c \cdot \Lambda(S, F) \ln(n), \tag{1}$$

where c is a numerical constant, S is the support of the signal, F is the distribution describing how to choose the blocks of measurements. In this bound, $\Lambda(S, F)$ is controlling the largest singular value of the sensing matrix restricted to the support of interest.

For a fixed signal x to be recovered, assuming that x has a random sign structure, we derive robust reconstruction guarantee with a required number of measurements:

$$m \geq c \cdot \Theta(S, F) \ln^2(n). \tag{2}$$

We study how far those CS results are from oracle-type guarantees: (i) $\Theta(S, F)$ is an upper bound of $\Lambda(S, F)$, (ii) there is an extra log factor. Actually by making an extra assumption, one can derive an oracle-type result in terms of number of

measurements as in (1). We also show that this additional assumption can be satisfied by very structured sampling matrices: for instance when sampling isolated measurements from a Fourier-wavelet transform (which can model Magnetic Resonance Imaging).

These results give an insight to design new optimal sampling strategies when realistic physical constraints are imposed in the acquisition. Indeed, one can minimize $\Theta(S, F)$ or $\Lambda(S, F)$ with respect to F given some prior on the support S. For instance in the case of sampling isolated measurements from an isometry, the state-of-the-art results [2, 3] consist in variable density sampling according to the probability distribution π, such that

$$\pi_k \propto \|a_k\|_\infty^2,$$

where the (a_k)'s are the measurement vectors. The new results show that one should instead sample according to the following probability distribution

$$\pi_k \propto \|a_k\|_\infty \|a_{k,S}\|_1.$$

This new strategy emphasizes that one should sample not only where the transform is coherent but also if there is information in the signal that can be captured (by some prior knowledge on the support S).

References

[1] C. Boyer, J. Bigot, P. Weiss *Compressed sensing with structured sparsity and structured acquisition*, ACHA (2017).
[2] N. Chauffert, P. Ciuciu, J. Kahn, P. Weiss *Variable density sampling with continous sampling trajectories*, SIAM Journal on Imaging Science (2014).
[3] F. Krahmer, R. Ward *Stable and robust sampling strategies for compressive imaging*, Image Processing, IEEE Transactions on, 23(2):612–622 (2014).

Stable Phase Retrieval in Infinite Dimensions

Rima Alaifari
(joint work with Ingrid Daubechies, Philipp Grohs, Rujie Yin)

The problem of phase retrieval originated from X-ray crystallography, in which an electron density distribution of a crystal or crystallized molecule is sought to be reconstructed from only the magnitude of its Fourier transform. In the more recent technique of coherent diffraction imaging for imaging of non–crystalline nanoscale structures, one way to retrieve lost phase information is by adding redundancy in the measurements. This is typically realized by a pinhole sliding over the object support. Such measurements can be modelled as the intensities of a short–time Fourier transform (STFT) of the underlying density. Another instance of a phase retrieval problem comes from the phase vocoder in audio processing. A phase vocoder is a device that realizes modifications of audio signals, such as time–stretching or pitch–shifting. One way to implement such modifications is through fitting an audio signal to a modified spectrogram, i.e. to STFT magnitudes.

In our work, we consider the problem of phase retrieval in an infinite–dimensional Hilbert space setting. More precisely, given a Hilbert space $\mathcal{H}$ and a frame $\{\psi_\lambda\}_{\lambda\in\Lambda} \subset \mathcal{H}$ for some index set Λ, we ask when a signal $f \in \mathcal{H}$ can be uniquely and stably determined from $\{|\langle f, \psi_\lambda\rangle|\}_{\lambda\in\Lambda}$ up to a global phase factor $\tau \in S^1$. By stability we mean the existence of uniform constants $c_1, c_2 > 0$ s.t. for all $f, g \in \mathcal{H}$,

$$c_1 \operatorname{dist}(f,g) \leq \|\{|\langle f, \psi_\lambda\rangle|\}_{\lambda\in\Lambda} - \{|\langle g, \psi_\lambda\rangle|\}_{\lambda\in\Lambda}\|_{L^2(\Lambda,\mu)} \leq c_2 \operatorname{dist}(f,g),$$

where $\operatorname{dist}(f,g) := \inf_{\tau\in S^1} \|f - \tau g\|_{\mathcal{H}}$.

Clearly, if the frame is not sufficiently redundant, phase retrieval is not uniquely solvable, because too much information has been lost. On the other hand, in certain examples, one can show that sufficient oversampling of a frame can restore the unique solvability of phase retrieval. For example, the reconstruction of real–valued signals $f \in L^2(\mathbb{R})$ from $\{|\langle f, \psi_\lambda\rangle|\}_{\lambda\in\Lambda}$ is possible when $\{\psi_\lambda\}_{\lambda\in\Lambda}$ is a Meyer wavelet frame obtained from oversampling a Meyer wavelet orthonormal basis by a factor of at least 16/3 [1].

A natural question that arises is whether oversampling can also be a tool for restoring the stability of phase retrieval. In [4], it has been proven that when $\mathcal{H}$ is infinite–dimensional and Λ is a discrete index set, phase retrieval can never be uniformly stable. We show in [3] that this is also the case when Λ is allowed to be a continuous index set. Thus, oversampling cannot improve the stability properties of phase recovery. More precisely, we prove a conjecture formulated in [5] for the finite–dimensional case, stating that the so–called *strong complement property (SCP)* is a necessary condition for stability of phase retrieval. Furthermore, we demonstrate that the SCP can never hold when $\mathcal{H}$ is infinite–dimensional.

Consequently, a function $f \in L^2(\mathbb{R})$ cannot even be stably recovered from the magnitudes of its continuous STFT or of its continuous wavelet transform, i.e. even in cases where the signal transform is not sampled at all. While this result on the stability of the problem is negative, we have noticed that in practice, the instabilities that occur are all of a certain kind. Whenever the signal transform is concentrated on at least two disjoint regions of the time–frequency/time–scale domain, and small outside of these regions, phase retrieval up to one global phase factor is no longer possible in practice.

On the positive side, we have made the observation that for audio signals that have such STFTs or wavelet transforms, multiplying the signal transform by a phase factor on only one of these regions results in an audio signal that is audibly identical to the original one (although they are no longer equal up to a *global* phase factor). More precisely, suppose that for example the STFT F of f is concentrated on two disjoint regions $D_1, D_2 \subset \mathbb{C}$, so that $F = F_1 + F_2$ and F_i is small outside of $D_i, i = 1, 2$. Then, the audio signal $\widetilde{f}$ for which the STFT is equal to $\widetilde{F} = F_1 + \tau F_2$, for $\tau \in S^1$, is audibly indistinguishable from f.

This observation has led us to formulate a new paradigm for stable phase retrieval in audio processing applications [2]. We propose to consider so–called *atoll functions*, i.e. functions concentrated on disjoint *atolls* and so that they are small

outside of these atolls. In the example from above, F is an atoll function concentrated on the two atolls D_1 and D_2. Then, if one aims to reconstruct up to a global phase factor on *each atoll separately,* stability can be restored. For this, the requirement on the atoll function is that it is bounded below on the atolls (here, one can allow small *lagoons* with possibly smaller values inside the atoll, where the sizes and the number of the lagoons will enter in the stability constant).

Our result holds for signal transforms that are holomorphic up to a weight function, i.e. for the STFT with Gaussian window and the continuous wavelet transform with Cauchy wavelet. An open question we believe to be interesting is that of extending this result to more general window classes and wavelets.

References

[1] R. Alaifari, I. Daubechies, P. Grohs, and G. Thakur, *Reconstructing real–valued functions from unsigned coefficients with respect to wavelet and other frames*, Journal of Fourier Analysis and Applications 23.6 (2017): 1480-1494.

[2] R. Alaifari, I. Daubechies, P. Grohs, and R. Yin, *Stable phase retrieval in infinite dimensions*, arXiv preprint arXiv:1609.00034 (2016).

[3] R. Alaifari, and P. Grohs, *Phase retrieval in the general setting of continuous frames for Banach spaces*, SIAM Journal on Mathematical Analysis 49.3 (2017): 1895-1911.

[4] J. Cahill, P. Casazza and I. Daubechies, *Phase retrieval in infinite–dimensional Hilbert spaces*, Transactions of the American Mathematical Society, Series B 3, (3) (2016), 63–76.

[5] A.S. Bandeira, J. Cahill, D.G. Mixon, and A.A. Nelson, *Saving phase: Injectivity and stability for phase retrieval*, Applied and Computational Harmonic Analysis 37.1 (2014): 106-125.

Low-rank Recovery from Group Orbits

David Gross

(joint work with Richard Kueng, Markus Grassl, Huangjun Zhu)

We are concerned with the problem of recovering an unknown $d \times d$ matrix X from m noisy rank-one measurements of the form

$$y_i = \mathrm{tr} X a_i a_i^* + \epsilon_i, \qquad i = 1, \dots, m$$

where $a_i \in \mathbb{C}^d$ are measurement vectors, and ϵ_i represents noise. We assume that a_i are sampled from a group orbit. More precisely, we fix some finite group $G \subset U(\mathbb{C}^d)$ and a "fiducial vector" $a \in \mathbb{C}^d$. The orbit is then $\mathcal{O} = \{ga \,|\, g \in G\}$. We assume that a (and hence all elements in the orbit) are normalized in that $\|a\|_2 = 1$. The $a_1, \dots, a_m$ are assumed the be sampled independently from $\mathcal{O}$.

The basic insight of Ref. [1] is that in this setting, recovery guarantees can sometimes be proven using just representation-theoretic data about G.

Our approach works like this: The basis are the results of Ref. [2] that establish low-rank recovery guarantees using just information about the fourth moments

$$M = \mathbb{E}[(a_i a_i^*)^{\otimes 4}] \tag{1}$$

of the rank-1 measurement matrices $a_i a_i^*$. Roughly speaking, the methods of Ref. [2] require that the matrix element

$$(b^{\otimes 4})^* M b^{\otimes 4} \tag{2}$$

be "small" for all normalized vectors $b \in \mathbb{C}^d$.

If the random vector a_i is sampled from a G-orbit, it has the same distribution as $g a_i$ for any $g \in G$. It follows that $g^{\otimes 4} M (g^{-1})^{\otimes 4} = M$ or, equivalently, $[M, g^{\otimes 4}] = 0$. We can thus apply Schur's Lemma. Assume for simplicity that all irreducible representations (irreps) of G that appear in the representation $g \mapsto g^{\otimes 4}$ are non-degenerate. In this case, Schur's Lemma says that

$$M = \sum_i \alpha_i P_i,$$

where i labels irreps, P_i projects on the i-th irrep, and the α_i are suitable coefficients.

From Eq. (1), one finds that $\mathrm{tr} M = 1$ and that M is positive semi-definite. Hence the coefficients fulfill $0 \leq \alpha_i \leq 1/\mathrm{tr} P_i$. Therefore, a sufficient condition for (2) to be small is that the dimensions of all irreps occuring in $g^{\otimes 4}$ is large. In fact, one can easily verify that one can restrict attention to irreps that are contained in the totally symmetric subspace $\mathrm{Sym}^4(\mathbb{C}^d) \subset (\mathbb{C}^d)^{\otimes 4}$. By the polarization identity, this space is equivalent to the space of degree-4 polynomials in d complex variables. *In this way, just using the existing techniques in* [2]*, we get stable uniform recovery guarantees for rank-1 measurements that are sampled from any orbit of any matrix group whose action on order-4 polynomials does not contain small irreps.*

In Refs. [1, 3], we use slightly strengthened arguments to show that a certain Clifford group satisfies these criteria. The Clifford group plays a central role in quantum information theory, and has long been studied e.g. in classical coding theory. Refinements of the representation-theoretica analysis and further applications appear in Refs. [4, 5].

References

[1] H. Zhu, R. Kueng, D. Gross, *Low rank matrix recovery from Clifford orbits*, arXiv:1610.08070.

[2] R. Kueng, H. Rauhut, and U. Terstiege, *Low rank matrix recovery from rank one measurements*, Appl. Comput. Harmonic Anal., 2015.

[3] H. Zhu, R. Kueng, M. Grassl, D. Gross, *The Clifford group fails gracefully to be a unitary 4-design*, arXiv:1609.08172.

[4] R. Kueng, H. Zhu, D. Gross, *Distinguishing quantum states using Clifford orbits*, arXiv:1609.08595.

[5] D. Gross, S. Nezami, M. Walter, *Schur-Weyl Duality for the Clifford Group with Applications*, arXiv:1712.08628.

On unlimited sampling

Felix Krahmer
(joint work with Ayush Bhandari, Ramesh Raskar)

For the conversion of an analog (bandlimited) signal to a digital representation, so called analog–to–digital converters (ADCs) are of key importance. The role of such devices is to extract from an analog signal its values on a discrete grid. Provided these samples are taken at a high enough rate, they then allow for the recovery of the signal via Shannon's sampling theorem. Unlike the sampling method assumed in Shannon's sampling theorem, practical ADCs are limited in dynamic range. Whenever a signal exceeds some preset threshold, the ADC saturates, resulting in aliasing due to clipping.

Recent developments in ADC design, allow for an alternative ADC construction, namely ADCs that reset rather than to saturate, thus producing modulo samples. Depending on the community, the resulting ADC constructions are known as *folding*-ADC (cf. [1] and references therein) or the *self-reset*-ADC, recently proposed by Rhee and Joo [2] in context of CMOS imagers. More precisely, when reaching the upper or lower saturation threshold $\pm\lambda$, these ADCs would reset to the respective other threshold, i.e., $\mp\lambda$, in this way allowing to capture subsequent changes even beyond the saturation limit. Mathematically, this is represented by a memoryless, non-linear mapping of the form

$$\mathcal{M}_\lambda : t \mapsto 2\lambda \left(\left[\!\!\left[\frac{t}{2\lambda} + \frac{1}{2} \right]\!\!\right] - \frac{1}{2} \right). \tag{1}$$

These constructions give rise to the following mathematical problem. Given such modulo samples of a bandlimited function as well as the dynamic range of the ADC, how can the original signal be recovered and what are the sufficient conditions that guarantee perfect recovery?

The following theorem provides such a sufficiency condition.

Theorem 1 (Unlimited Sampling Theorem [3])**.** *Let $g(t)$ be π-bandlimited and consider, for $k \in \mathbb{Z}$, the modulo samples $y_k = \mathcal{M}_\lambda(g(kT))$ of $g(t)$ with sampling rate T. Then a sufficient condition for recovery of $g(t)$ from the $\{y_k\}_k$ up to additive multiples of 2λ is that*

$$T \leq \frac{1}{2\pi e}. \tag{2}$$

At the core of Theorem 1 is a constructive recovery method, as summarized in Algorithm . While some estimate of the signal norm needs to be available when recovering, the underlying circuit architecture is not limited to certain amplitude ranges: the same architecture allows for the recovery of arbitrary large amplitudes. That is why we refer to our approach as unlimited sampling.

The underlying observation of our recovery algorithm is that for significant oversampling, the size of the n-th order finite difference scales like the n-th power

Algorithm 1 Recovery from Modulo Folded Samples

Data: $y_k = \mathcal{M}_\lambda(g(kT))$, $N \in \mathbb{N}$, and $2\lambda\mathbb{Z} \ni \beta_g \geq \|g\|_\infty$.
Result: $\widetilde{g} \approx g$.

(1) Compute $\overline{y} = \Delta^N y$.
(2) Compute $\overline{\varepsilon}_\gamma = \mathcal{M}_\lambda(\overline{y}) - \overline{y}$. Set $s_{(1)} = \overline{\varepsilon}_\gamma$.
(3) for $n = 1 : N-1$
  Compute $\kappa_{(n)}$ in (4).
  $s_{(n+1)} = \mathsf{S}s_{(n)} - 2\lambda\kappa_{(n)}$.
  end
(4) $\widetilde{\gamma} = \mathsf{S}s_{(N)}$.
(5) Compute $\widetilde{g}$ from $\widetilde{\gamma}$ via low-pass filter.

of the oversampling rate and hence becomes small. In addition, the finite difference operator and the modulo operation $\mathcal{M}_\lambda$ satisfy the following commutativity relation.

Proposition 1. *For any sequence a it holds that*

$$\mathcal{M}_\lambda(\Delta^N a) = \mathcal{M}_\lambda(\Delta^N(\mathcal{M}_\lambda(a)). \tag{3}$$

Combining these observations allows for the recovery of the finite differences. Namely, the right hand side of (3) can be computed from the modulo samples, which hence provides access to the left hand side. As the argument of the modulo operation on the left hand side is small, the operation has no effect, so one has computed the true finite difference.

To invert the finite difference operation, we consider the difference between true samples and modulo samples, which will always lie on a grid of spacing 2λ. For this reason, the inversion will be considerably more stable than for arbitrary real inputs. In particular, the integration constant that introduces ambiguity in each of the inversion steps will also lie on a grid. As a consequence, choosing the wrong constant will cause the output function in the subsequent step to exhibit a very strong growth, which in turn can be detected when using enough samples. In this estimate, the a priori bound of the amplitude of the signal $2\lambda\mathbb{Z} \ni \beta_g \geq \|g\|_\infty$ plays an important role. Namely, for $J = \frac{6\beta_g}{\lambda}$, the appropriate inverse of the n-th finite difference operator is given by the sequence of partial sums (this operation is denoted by S) adjusted by a constant of $2\lambda\kappa_n$, where

$$\kappa_{(n)} = \left\lfloor \frac{(\mathsf{S}^2\Delta^n\varepsilon_\gamma)_1 - (\mathsf{S}^2\Delta^n\varepsilon_\gamma)_{J+1}}{8\beta_g} + \frac{1}{2} \right\rfloor. \tag{4}$$

When the bandlimited signals under consideration have additional structure, a corresponding approach can sometimes allow the recovery from just finitely many modulo samples (e.g., in the context of superresolution [4] or for sums of sinusoids [5]). In more general scenarios without smoothness assumptions, such as

redundant representations in $\mathbb{R}^N$, it is not clear under which conditions comparable recovery guarantees can be obtained. We consider this to be an interesting follow-up problem.

References

[1] W. Kester, *ADC architectures VI: Folding ADCs (MT-025 tutorial)* Analog Devices, Tech. Rep., 2009.

[2] J. Rhee and Y. Joo, *Wide dynamic range CMOS image sensor with pixel level ADC,* Electron. Lett., **39** (4), 360, 2003.

[3] A. Bhandari, F. Krahmer, R. Raskar, *On Unlimited Sampling*, Intl. Conf. Sampling Theory Appl. (SampTA), 2017

[4] A. Bhandari, F. Krahmer and R. Raskar, *Unlimited Sampling of Sparse Signals,* to appear in IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[5] A. Bhandari, F. Krahmer and R. Raskar, *Unlimited Sampling of Sparse Sinusoidal Mixtures,* to appear in IEEE Intl. Symp. on Information Theory (ISIT), 2018.

Twisted X-rays – mathematical design of radiation for high-resolution X-ray diffraction imaging

Dominik Juestel

(joint work with Gero Friesecke, Richard D. James)

Conventional methods for X-ray diffraction imaging use plane waves to illuminate a sample. The incoming electromagnetic radiation induces oscillations of the electron density of the sample. Consequently, the moving charges produce an outgoing field that can be recorded by detectors. The inverse problem to infer the structure of the sample form its diffraction patterns can be formulated as a phase retrieval problem: only the absolute values of the complex numbers that are needed for reconstruction can directly be obtained from the measured diffraction intensities. More precisely, the information about the electron density ρ that is contained in the measurements is essentially the modulus of its Fourier transform $|\widehat{\rho}|$.

Atomic resolution X-ray diffraction imaging can today only be achieved by X-ray crystallography, where the periodic arrangement of molecules leads to a high amount of constructive and destructive interference, resulting in highly structured peak patterns. Mathematically, this effect is explained by the Poisson summation formula. Let $\Gamma := \mathbf{A}\mathbb{Z}^3$, $\mathbf{A} \in \mathrm{GL}(3,\mathbb{R})$, be a periodic lattice in $\mathbb{R}^3$, then a crystal's electron density can be modeled as an infinite periodic function $\rho = \delta_\Gamma * \varphi$, where $\delta_\Gamma := \sum_{x\in\Gamma} \delta_x$ is the Dirac comb of the lattice Γ, and φ is a model for the electron density in a unit cell. Combining the Poisson summation formula with the convolution theorem, we get $|\widehat{\rho}| = \frac{(2\pi)^3}{\det(\mathbf{A})}\delta_{\Gamma'} \cdot |\widehat{\varphi}|$, where $\Gamma' := 2\pi\mathbf{A}^{-T}\mathbb{Z}^3$ is the corresponding reciprocal lattice. This calculation shows, that the intensity measured at a diffraction peak is essentially the modulus of a Fourier coefficient of the electron density in a unit cell of the crystal.

The main drawback of X-ray crystallography is the need to crystallize the structures under consideration. Since proteins, for example, often do not form crystals,

but aggregate in other highly symmetric assemblies, like rods, sheets, or icosahedral structures, our approach is to design forms of electromagnetic radiation that reflect the symmetry of the considered class of structures in the same way as plane waves reflect the symmetry of crystals. This way, we can profit from the inteference effects while the structures keep their natural form. Mathematically, this is achieved by finding the eigenfunctions of the group action of the symmetry group in the space of monochromatic solutions to Maxwell's equations in vacuum. More general versions of the Poisson summation formula then imply highly structured interference similar to the classic case (see [2]).

Unlike in classic X-ray crystallography, where the vectorial nature of the electromagnetic field can be neglected in most calculations, it needs to be taken into account for more general radiation than plane waves. When translating a vector field, the orientation of the field vectors doesn't change. Instead, when rotating or reflecting a vector field, the field vectors need to be rotated or reflected accordingly. This simple and intuitive fact has implications for the reconstruction problem from non-plane wave diffraction patterns. The classic scalar phase retrieval problem is generalized to a vectorial phase problem: the information contained in the intensity measurements is the length of complex vectors that are needed for reconstruction of the electron density.

In the special case of helical structures like nanotubes or helical viruses – these are structures that have a discrete symmetry of rotations, translations, and screw displacements with respect to a fixed axis – the radiation design problem can be fully solved. We call the resulting electromagnetic fields twisted X-rays, as they are waves that propagate along helices. The solution spaces are finite dimensional, with a parametrization that can be interpreted as the polarization of the radiation (see [3]).

A diffraction experiment of twisted X-rays illuminating an aligned helical structure is conjectured to produce a highly structured diffraction pattern, with the outgoing field in axial direction forming an exact double peak pattern when viewed as a function of certain radiation parameters. Moreover, the structure of the sample can be recovered by solving a variation of the classic phase retrieval problem. In fact, the above mentioned vectorial phase retrieval problem reduces to a scalar phase retrieval problem, with the Fourier transform replaced by a Fourier-Hankel transform.

For general symmetries, the analysis of the solution spaces of the design equations gets more involved, and can in general not be made as explicit as in the case of plane or twisted waves. Viewing the design problem from the standpoint of representation theory, the space of monochromatic solutions to Maxwell's equations in vacuum is decomposed into irreducible components with respect to the action of the symmetry group. These components need not be finite-dimensional, as is the case for the translation group or the helical symmetry group.

When trying to calculate the diffraction patterns that result from the illumination with radiation from an irreducible component with respect to a general symmetry group, one arrives at the boundary of mathematical research. While

for abelian and compact groups, there are generalizations of the classic theory, for non-abelian non-compact groups, there is no suitable generalization of the Poisson summation formula (see [4] for the mathematical background).

Recent advances in X-ray technology suggest that the proposed experiment, to illuminate helical structures with twisted X-rays, might be realizable in the near future. Several groups managed to produce so-called beams carrying orbital angular momentum, which are closely related to twisted X-rays (see, e.g., [1]). They use helical undulators to force an electron beam from a synchrotrone onto a helical trajectory. The X-rays that are emitted interfere to form the helical waveform. While they do not yet reach the energy that is necessary for atomic resolution, a proof of concepts is within reach. The proposed method has the potential to allow for structure analysis of previously unaccessible molecules, while the theory is a nice example for the usefullness of abstract mathematics in applications.

References

[1] J. Bahrdt, K. Holldack, P. Kuske, R. Müller, M. Scheer, P. Schmid, *First observation of photons carrying orbital angular momentum in undulator radiation*, Phys. Rev. Lett. **111** (2013), 034801.

[2] G. Friesecke, R. D. James, D. Juestel, *Twisted X-rays: incoming waveforms yielding discrete diffraction patterns for helical structures*, SIAM J. Appl. Math. **76-3** (2016), 1191–1218.

[3] D. Juestel, G. Friesecke, R. D. James, *Bragg-Von Laue diffraction generalized to twisted X-rays*, Acta Cryst. A **72** (2016), 190–196.

[4] D. Juestel, *The Zak transform on strongly proper G-spaces and its applications*, J. London Math. Soc. **97 (2)** (2018), 47–76.Calderbank

Synchronization Problems: Geometry Meets Learning

Tingran Gao

(joint work with Ingrid Daubechies, Sayan Mukherjee, Doug Boyer, Jacek Brodzki, Qixing Huang, Chandrajit Bajaj)

Acquiring complex, massive, and often high-dimensional data sets has become a common practice in many fields of science. Bridging recent developments applying differential geometry and topology in probability and statistical sciences, the problem of *synchronization* arise in a variety of fields in computer vision, signal processing, combinatorial optimization, and natural sciences (e.g. cryoelectron microscopy and geometric morphometrics [3, 4]). The data given in a synchronization problem include a connected graph that encodes similarity relations within a collection of objects, and pairwise correspondences—often realized as elements of a transformation group G—characterizing the nature of the similarity between a pair of objects linked directly by an edge in the relation graph. The goal is to adjust the pairwise correspondences, which often suffer from noisy or incomplete measurements, to obtain a globally consistent characterization of the pairwise relations for the entire dataset, in the sense that unveiling the transformation between

a pair of objects far-apart in the relation graph can be done by composing transformations along consecutive edges on a path connecting the two objects, and the resulting composed transformation is independent of the choice of the path.

We develop a geometric framework in [1] that characterizes the nature of synchronization based on the classical theory of fibre bundles. We first establish the correspondence between synchronization problems in a topological group G over a connected graph Γ and the moduli space of flat principal G-bundles over Γ, and develop a discrete analogy of the renowned theorem of classifying flat principal bundles with fixed base and structural group using the representation variety. In particular, we show that prescribing an edge potential on a graph is equivalent to specifying an equivalence class of flat principal bundles, of which the triviality of holonomy dictates the synchronizability of the edge potential.

Based on the fibre bundle interpretation of synchronization problems, we develop in [1] a twisted cohomology theory for associated vector bundles of the flat principal bundle arising from an edge potential, which is a discrete version of the twisted cohomology in differential geometry. This leads to a twisted Hodge theory, which is a fibre bundle analog of the discrete Hodge theory on graphs. The lowest-degree Hodge Laplacian of this twisted Hodge theory recovers a geometric realization of the graph connection Laplacian (GCL), a group-valued graph operator studied extensively in synchronization problems. Similar intuitions have led to an extended diffusion geometry framework for datasets with an underlying fibre bundle structure, referred to as *Horizontal Diffusion Maps* [2], which models a dataset with pairwise structural correspondences as a fibre bundle equipped with a connection; the role of random walk in standard diffusion maps is replaced with a horizontal random walk on the fibre bundle driven by a random walk on the base space. This novel diffusion geometry framework demonstrates its advantage of leveraging more detailed structural information to improve clustering accuracy in automated geometric morphometrics [5].

The geometric framework established in [1] also motivated us to study the problem of learning group actions—partitioning a collection of objects based on the local synchronizability of pairwise correspondence relations. A dual interpretation is to learn finitely generated subgroups of an ambient transformation group from noisy observed group elements. An iterative two-step synchronization residual spectral clustering algorithm is proposed in [1]. More concretely, assuming the underlying graph consists of multiple clusters, and the transformation groups within each cluster is more consistent than between clusters, the algorithm performs a synchronization procedure over the entire graph, followed by evaluating the discrepancy ("edgewise frustration") between the synchronized and the original edge potentials, and then performs a spectral clustering for the graph with the edgewise frustration as weights; after that, the algorithm runs synchronization within each cluster, patches the local synchronization solutions together, and repeat the steps starting from another global synchronization. This simple algorithm demonstrates its efficacy on both simulated and real datasets. When the group is

a permutation group, we established in [6] exact recovery conditions for this iterative synchronization-residual based clustering algorithm under a stochastic block model nested with inhomogeneous random corruption.

Many exciting problems are still open in this line of research. Notably, a proper analogy of the Cheeger inequality seems natural but is missing so far in the principal bundle framework. Much more about the horizontal diffusion geometry is unknown either, for instance, whether the eigenfunctions of the horizontal diffusion operator can be manipulated to obtain an embedding of either the total space or the base space of the fibre bundle. It is also of great interest to generalize the techniques developed in [6] to establish similar results for groups other than the permutations group, such as orthogonal or special orthogonal groups, which are commonly encountered in shape alignment and analysis. Last but not the least, we are excited about the connection between differential geometry and learning theory implied by our geometric framework, which seems to suggest that the interchanging of ideas from either field can substantially benefit the other.

References

[1] T. Gao, J. Brodzki, S. Mukherjee, *The Geometry of Synchronization Problems and Learning Group Actions*, submitted. arXiv:1610.09051, (2016)

[2] T. Gao, *The Diffusion Geometry of Fibre Bundles*, submitted. arXiv:1602.02330, (2016).

[3] T. Gao, G.S. Yapuncich, I. Daubechies, S. Mukherjee, and D.M. Boyer, *Development and Assessment of Fully Automated and Globally Transitive Geometric Morphometric Methods, with Application to a Biological Comparative Dataset with High Interspecific Variation*, The Anatomical Record. **to appear**. (2017).

[4] N.S. Vitek, C.L. Manz, T. Gao, J.I. Bloch, S.G. Strait, and D.M. Boyer, *Semi-Supervised Determination of Pseudocryptic Morphotypes Using Observer-Free Characterizations of Anatomical Alignment and Shape*, Methods in Ecology and Evolution, **7** (2017) 5041-5055.

[5] T. Gao, *Hypoelliptic Diffusion Maps and Their Applications in Automated Geometric Morphometrics*, PhD thesis, Duke University. (2015)

[6] C. Bajaj, T. Gao, Z. He, Q. Huang, and Z. Liang, *SMAC: Simultaneous Mapping and Clustering Using Spectral Decompositions*, submitted to 2018 International Conference on Machine Learning. (2018)

Solving nonlinear equations using convex programming

Justin Romberg

(joint work with Sohail Bahmani)

We consider the general problem of recovering an unknown vector $\boldsymbol{x}_\star \in \mathbb{R}^N$ that (approximately) satisfies a system of equations

$$\begin{aligned} y_1 &= f_1(\boldsymbol{x}_\star) + \epsilon_1 \\ y_2 &= f_2(\boldsymbol{x}_\star) + \epsilon_2 \\ &\vdots \\ y_M &= f_M(\boldsymbol{x}_\star) + \epsilon_M, \end{aligned}$$

where the f_m are known, convex functions and the ϵ_m are unknown perturbations.

We attack this problem as follows. Setting the perturbations $\epsilon_m = 0$ for now to ease the explanation, each equation above gives us a different *convex* feasibility region for $\boldsymbol{x}_\star$, namely the sublevel set $\{\boldsymbol{x} \ : \ f_m(\boldsymbol{x}) \leq y_m\}$. It is clear $\boldsymbol{x}_\star$ must lie in the intersection of these sublevel sets:

$$\boldsymbol{x}_\star \in \mathcal{K}, \quad \mathcal{K} = \bigcap_{m=1}^{M} \{\boldsymbol{x} \ : \ f_m(\boldsymbol{x}) \leq y_m\}.$$

In fact, $\boldsymbol{x}_\star$ must be an extreme point of $\mathcal{K}$. As such, we can attempt to recover $\boldsymbol{x}_\star$ by maximizing a linear functional over $\mathcal{K}$. For a given $\boldsymbol{a}_0$, we solve

$$\underset{\boldsymbol{x}}{\text{maximize}} \ \langle \boldsymbol{x}, \boldsymbol{a}_0 \rangle \quad \text{subject to} \quad \boldsymbol{x} \in \mathcal{K}. \tag{1}$$

It is clear that if $\boldsymbol{a}_0 = \boldsymbol{x}_\star$, then $\boldsymbol{x}_\star$ is the solution (or at least one of the solutions) to the program above. Our work develops conditions under which there are *many* $\boldsymbol{a}_0$ such that $\boldsymbol{x}_\star$ is the unique solution to (1). In particular, we assume that we have a vector $\boldsymbol{a}_0$ that is only roughly correlated with $\boldsymbol{x}_\star$:

$$\frac{\langle \boldsymbol{x}_\star, \boldsymbol{a}_0 \rangle}{\|\boldsymbol{x}_\star\|_2 \|\boldsymbol{a}_0\|_2} \ \geq \ \delta > 0, \tag{2}$$

for some constant δ. We call such a $\boldsymbol{a}_0$ an *anchor vector*.

Simply writing down the optimality (KKT) conditions for (1) shows us that if indeed $y_m = f_m(\boldsymbol{x}_\star)$, then $\boldsymbol{x}_\star$ is a solution to (1) if and only if

$$\boldsymbol{a}_0 \in \operatorname{cone}\left(\{\nabla f_m(\boldsymbol{x}_\star), \ m = 1, \ldots, M\}\right). \tag{3}$$

This tells us that whether or not $\boldsymbol{a}_0$ will be effective is purely a function of the behavior of the gradients of the f_m at the solution. Qualitatively, we can see that the more diverse these gradients are, the larger the cone that they generate is going to be, and the easier it is to find a suitable $\boldsymbol{a}_0$.

Our main result gives a guarantee on the number of equations we need for (3) to hold with high probability when the functions are drawn independently and identically distributed according to some probability law. Under this probability law, we let $\boldsymbol{\Sigma}_\star$ be the correlation matrix of the gradients at the solution

$$\boldsymbol{\Sigma}_\star = \mathbb{E}[\nabla f(\boldsymbol{x}_\star) \nabla f(\boldsymbol{x}_\star)^{\mathrm{T}}]$$

and define the quantities

$$\tau_\star = \inf_{\|\boldsymbol{h}\|_2 = 1} \mathbb{E}[\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle_+], \quad \nu_\star = \frac{\|\boldsymbol{\Sigma}_\star\|}{\tau_\star},$$

where $\langle \cdot, \cdot \rangle_+$ takes the positive part of the inner product. Then if we have an $\boldsymbol{a}_0$ that obeys (2), $\boldsymbol{x}_\star$ will be the solution to (1) with high probability when

$$M \ \geq \ \text{Const} \cdot \nu_\star^3 \cdot N.$$

When the $\nabla f(\boldsymbol{x}_\star)$ is approximately "isotropic", the quantity $\nu_\star$ will be a constant, and the system of equations $y_m = f_m(\boldsymbol{x}_\star)$ can be solved for $M \sim N$.

We can also encourage certain types of structure in the solution by adding a regularizer to (1). For example, it is now well-known that sparsity in the solution

to a system of equations can be encouraged by penalizing with an ℓ_1 norm. In general, if $\Omega(\boldsymbol{x})$ is a convex regularizer, then we can solve

$$\underset{\boldsymbol{x}}{\text{maximize}} \ \langle \boldsymbol{x}, \boldsymbol{a}_0 \rangle - \Omega(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{x} \in \mathcal{K}. \tag{4}$$

In this case, if $y = f_m(\boldsymbol{x}_\star)$, then $\boldsymbol{x}_\star$ is the solution to this optimization program if and only if

$$\boldsymbol{a}_0 \in \text{cone}\left(\{\nabla f_m(\boldsymbol{x}_\star),\ m = 1, \dots, M\}\right) + \partial\Omega(\boldsymbol{x}_\star),$$

where $\partial\Omega(\boldsymbol{x}_\star)$ is the subgradient of $\Omega(\cdot)$ at $\boldsymbol{x}_\star$. Notice that if $\boldsymbol{x}_\star$ lies on a "corner" of the sublevel set $\{\boldsymbol{x} : \Omega(\boldsymbol{x}) \leq \Omega(\boldsymbol{x}_\star)\}$, then $\partial\Omega(\boldsymbol{x}_\star)$ can be large, allowing for a more liberal choice of anchor vector $\boldsymbol{a}_0$.

For random f_m, the number of equations we need for $\boldsymbol{x}_\star$ to be the solution to (4) again depends on a measure of statistical complexity of the gradients $\nabla f(\boldsymbol{x}_\star)$, but now relative to the ascent cone for the functional in (4). An exact statement of this result can be found in [2].

References

[1] S. Bahmani and J. Romberg, *A flexible convex relaxation for phase retrieval*, Electronic Journal of Statistics, vol. 11, no. 2 (2017), 5254–5281.

[2] S. Bahmani and J. Romberg, *Solving equations of random convex functions via anchored regression*, Preprint, September 2017, arxiv:1702.05327.

Fast Point Cloud Distances and Multi-Sample Testing

Alex Cloninger

(joint work with Xiuyuan Cheng and Ronald R. Coifman)

We consider the question of estimating the total variation between two distributions in high dimensional space, given only a finite number of samples drawn iid from each distribution. This talk introduces a new anisotropic kernel-based Maximum Mean Discrepancy (MMD) statistic for estimating such a distance [1], which builds upon the Reproducing Kernel Hilbert Space MMD proposed by Gretton, et al [2]. The new anisotropic kernel scales linearly in the number of points by establishing landmarks throughout the space that approximate the local geometry of the union of the two datasets by constructing principle components of local covariance matrices. These landmarks can be interpreted as points that, under the action of a heat kernel on the data, diffuse to the entire space as quickly as possible. When the distributions are locally low-dimensional, the proposed test can be made more powerful to distinguish certain alternatives. While the proposed statistic can be viewed as a special class of Reproducing Kernel Hilbert Space MMD, the consistency and power of the test is proved, under mild assumptions of the kernel, as long as $\|p - q\| \sim \mathbb{O}(n^{-1/2+\delta})$ for any $\delta > 0$, based on a result of convergence in distribution of the test statistic. We also establish error bounds on the approximation by landmark points.

We also consider the k-sample setting in which we measure pairwise distances between the k different point clouds. This test has complexity by $\mathbb{O}(\binom{k}{2}|R| + kN|R|d)$ for N points per cloud with $|R|$ landmarks in d dimensions. This is opposed to complexity $\mathbb{O}(\binom{k}{2}N^2d)$ of the naive algorithm of directly computing the MMD between any two distributions. Applications to flow cytometry detection of AML and diffusion MRI datasets are demonstrated, which motivate the proposed approach to compare distributions.

References

[1] X. Cheng, A. Cloninger, & R. R. Coifman, *Two-sample Statistics Based on Anisotropic Kernels*, arXiv preprint arXiv:1709.05006 (2017).

[2] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schlkopf, & A. J. Smola, *A kernel two-sample test*, Journal of Machine Learning Research **13** (2012), 723-773.

The Mismatch Principle: An Ignorant Approach to Non-Linear Compressed Sensing?

Martin Genzel

(joint work with Gitta Kutyniok, Peter Jung)

In many real-world problems, one is given a finite collection of samples

$$(\boldsymbol{a}_1, y_1), \dots, (\boldsymbol{a}_m, y_m) \in \mathbb{R}^n \times \mathbb{R}$$

which are drawn independently from a joint random pair $(\boldsymbol{a}, y)$ in $\mathbb{R}^n \times \mathbb{R}$ of unknown probability distribution. For example, $y \in \mathbb{R}$ could play the role of an *output variable* that one would like to predict from certain *input data* $\boldsymbol{a} \in \mathbb{R}^n$. Very generally speaking, the problem issue is now as follows:

> *What can we learn from the sample set about the relationship between the input and the output variables?*

Although we do not impose any specific restrictions on the model, it is useful to think of some (unknown) parameters that determine the underlying observation rule. Let us consider two prototypical scenarios:

- *Single-index models.* Let $\boldsymbol{x}_0 \in \mathbb{R}^n$ be a structured vector (e.g., sparse) and assume that
$$y = f(\langle \boldsymbol{a}, \boldsymbol{x}_0 \rangle)$$
where $f\colon \mathbb{R} \to \mathbb{R}$ can be unknown, non-linear, and noisy. The goal is to estimate the unknown index vector $\boldsymbol{x}_0$.
- *Variable selection.* Let $S = \{j_1, \dots, j_s\} \subset [n]$ and assume that
$$y_i = F(a_{j_1}, \dots, a_{j_s})$$
where $F\colon \mathbb{R}^s \to \mathbb{R}$ can be again unknown, non-linear, and noisy. The goal is to identify the set of active variables S in $\boldsymbol{a} = (a_1, \dots, a_n)$.

We would like to investigate whether a standard estimator — which does not require any prior knowledge — can solve those types of estimation problems. One of the most popular algorithmic approaches is the *generalized Lasso*,

$$(P_K) \qquad \min_{\boldsymbol{x}\in\mathbb{R}^n} \tfrac{1}{2m}\sum_{i=1}^{m}(y_i - \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle)^2 \quad \text{subject to} \quad \boldsymbol{x}\in K,$$

where $K \subset \mathbb{R}^n$ is a convex *hypothesis set* that enforces certain structural constraints on the solution, such as sparsity. The results of [1, 2, 3, 5] show that the Lasso — although originally designed for linear regression — is surprisingly robust against non-linear distortions and can in fact deal with much more complicated observation schemes. Let us state a simplified recovery guarantee:

Theorem 1 (informal, cf. [3, Thm. 6.4]) *Using the above notation, assume that* $\boldsymbol{a}$ *is an isotropic, mean-zero sub-Gaussian random vector in* $\mathbb{R}^n$ *and* y *is also sub-Gaussian. Fix an arbitrary* target vector $\boldsymbol{x}^\natural \in K \subset \mathbb{R}^n$. *Then, with high probability, any minimizer* $\hat{\boldsymbol{x}}$ *of* (P_K) *satisfies the following error bound:*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}^\natural\|_2 \lesssim \frac{w_\wedge(K, \boldsymbol{x}^\natural)}{\sqrt{m}} + \rho(\boldsymbol{x}^\natural), \tag{1}$$

where $w_\wedge(K, \boldsymbol{x}^\natural)$ *denotes the* conic Gaussian width *of* K *at* $\boldsymbol{x}^\natural$ *and*

$$\rho(\boldsymbol{x}^\natural) := \big\|\mathbb{E}[(\langle \boldsymbol{a}, \boldsymbol{x}^\natural\rangle - y)\boldsymbol{a}]\big\|_2$$

is called the mismatch covariance.

Remarkably, the above statement holds true for *every* choice of $\boldsymbol{x}^\natural$ and there are no specific assumptions on the output variable y. But in order to turn (1) into a meaningful error bound, one clearly needs to ensure that the offset term $\rho(\boldsymbol{x}^\natural)$ is sufficiently small, since it does not decay with m. If the target vector $\boldsymbol{x}^\natural$ can be chosen in this way, Theorem 1 states that the Lasso (P_K) indeed constitutes an almost consistent estimator of $\boldsymbol{x}^\natural$. With regard to our initial problem issue, we can now formulate a simplified version of the *mismatch principle*:

> *Determine a target vector* $\boldsymbol{x}^\natural \in K$ *that captures the "parametric" structure of the observation rule and minimizes the mismatch covariance* $\rho(\boldsymbol{x}^\natural)$ *at the same time.*

For example, we would have to specify a target vector in $\operatorname{span}\{\boldsymbol{x}_0\} \cap K$ for single-index models and in $\{\boldsymbol{x} \mid \operatorname{supp}(\boldsymbol{x}) \subseteq S\} \cap K$ for variable selection, respectively. It is in fact not hard to see that in either case (if $\boldsymbol{a}$ is standard Gaussian) there exists an appropriate choice of $\boldsymbol{x}^\natural$ such that $\rho(\boldsymbol{x}^\natural) = 0$.

In general, the mismatch principle provides a recipe to prove theoretical error bounds for the Lasso under non-linear observations. Combined with Theorem 1, it particularly indicates when one can expect reasonable outcomes and when not. A crucial role is obviously played by the mismatch covariance because it measures the compatibility between the linear fit of (P_K) and the true (parametric) model.

Apart from that, the mismatch principle even applies to more complicated situations, e.g., if the components of $\boldsymbol{a}$ are strongly correlated [4] or if the square loss in (P_K) is replaced by a different convex loss function [1].

References

[1] M. Genzel, *High-Dimensional Estimation of Structured Signals From Non-Linear Observations With General Convex Loss Functions*, IEEE Trans. Inf. Theory **63.3** (2017), 1601–1619.

[2] M. Genzel and P. Jung, *Blind Sparse Recovery From Superimposed Non-Linear Sensor Measurements*, In Proceedings of the 12th International Conference on Sampling Theory and Applications (SampTA), 2017.

[3] M. Genzel and P. Jung, *Recovering Structured Data From Superimposed Non-Linear Measurements*, Preprint arXiv:1708.07451, 2017.

[4] M. Genzel and G. Kutyniok, *A Mathematical Framework for Feature Selection from Real-World Data with Non-Linear Observations*, Preprint arXiv:1608.08852, 2016.

[5] Y. Plan and R. Vershynin, *The generalized Lasso with non-linear observations*, IEEE Trans. Inf. Theory **62.3** (2016), 1528–1537.

Dictionary learning - from local towards global and adaptive

Karin Schnass

The goal of dictionary learning is to decompose a data matrix $Y = (y_1, \dots, y_N)$, where $y_n \in \mathbb{R}^d$, into a dictionary matrix $\Phi = (\varphi_1, \dots, \varphi_K)$, where each column also referred to as atom is normalised, $\|\varphi_k\|_2 = 1$ and a sparse coefficient matrix $X = (x_1, \dots, x_N)$,

$$Y \approx \Phi X \tag{1}$$

One way to concretise that the coefficient matrix should be sparse is to choose a sparsity level S and ask that every coefficient vector x_n should have at most S non zero entries. Defining $\mathcal{D}_K$ to be the set of all dictionaries with K atoms and $\mathcal{X}_S$ the set of all columnwise S-sparse coefficient matrices the dictionary learning problem can be formulated as optimisation programme

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2. \tag{2}$$

This problem is highly non-convex and as such difficult to solve. However, randomly initialised alternating projection algorithms, which iterate between finding the best dictionary Ψ based on coefficients X and (trying) to find the best coefficients X based on a dictionary Ψ, such as K-SVD (K Singular Value Decompositions), [2], and ITKrM (Iterative thresholding and K residual means), [5], tend to be very successful on synthetic data - usually recovering 90 to 100% of all atoms - and provide useful dictionaries on image data.

The drawback of these algorithms is that assuming that the data Y is synthesized from a generating dictionary Φ and randomly drawn S- sparse coefficients X is that they have no (K-SVD) or very weak (ITKrM) recovery guarantees. This is in sharp contrast to more involved algorithms, which have gobal recovery guarantees but due to their computational complexity can only be used in toy examples in

small dimensions, [3, 1, 4].
One interesting exception is an algorithm developed by Sun, Qu and Wright, [7, 8], which is gradient based with a Newton trust region method to escape saddle points and proven to recover the generating dictionary if it is a basis. This result together with several results in machine learning which prove that non-convex problems can be well behaved, meaning all local minima are global minima, gives rise to hope that a similar result can be proven for learning overcomplete dictionaries.
In this talk and the accompanying paper, [6], we show that ITKrM has a much larger contraction radius, when assuming that the current estimate of the generating dictionary Ψ is incoherent and well conditioned. Assuming additionally that (after potential rearrangement and sign flip of the atoms) the cross-Gram matrix $\Psi^\star\Phi$ is diagonally dominant we further show that ITKrM is a contraction on as soon as

$$\max_k \|\varphi_k - \psi_k\|_2^2 \leq 2 - 2\|\Phi\|_{2,2} \log K \sqrt{S/K}, \tag{3}$$

which is relatively close to the worst case distance $\max_k \|\varphi_k - \psi_k\|_2^2 = 2$.
We also destroy all hope of proving global convergence by sketching the existence of stable fixed points, which are not equivalent to the generating dictionary. However, based on a characterisation of the fixed points and an analysis of the residuals at these fixed points we consider a replacement procedure for coherent atoms and develop a strategy for finding good replacement candidates. Decoupling the replacement strategy into independent pruning of coherent (as well as unused) atoms and adding of promising candidates finally leads to an algorithm for dictionary learning that adaptively chooses the dictionary size K and the sparsity level S.

References

[1] A. Agarwal, A. Anandkumar, and P. Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. In *COLT 2014 (arXiv:1309.1952)*, 2014.
[2] M. Aharon, M. Elad, and A.M. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing.*, 54(11):4311–4322, November 2006.
[3] S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *COLT 2014 (arXiv:1308.6273)*, 2014.
[4] B. Barak, J.A. Kelner, and D. Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *STOC 2015 (arXiv:1407.1543)*, 2015.
[5] K. Schnass. Convergence radius and sample complexity of ITKM algorithms for dictionary learning. *Applied and Computational Harmonic Analysis*, online, 2016.
[6] K. Schnass. Dictionary learning - from local towards global and adaptive. *arXiv:1804.07101*, 2018.
[7] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I: Overview and geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.
[8] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–915, 2017.

Monte Carlo approximation certificates for k-means clustering

Dustin G. Mixon
(joint work with Soledad Villar)

Geometric clustering is a fundamental problem in data science. At its core, clustering is an optimization problem, and it is natural to analyze the performance of various clustering routines with popular random data models. The last decade of research in applied harmonic analysis has demonstrated a fruitful interaction between optimization and randomness, suggesting that geometric clustering presents an opportunity for these techniques to contribute to the study of fundamental algorithms in data science. This talk discusses an instance of this opportunity and sheds light on additional examples that warrant further attention.

Given a finite sequence of data points $\{x_i\}_{i\in T}$ in $\mathbb{R}^m$ and a complexity parameter k, the k-means problem seeks a partition $C_1 \sqcup \cdots \sqcup C_k = T$ that minimizes the k-means objective:

(T-IP)

$$\text{minimize} \quad \frac{1}{|T|}\sum_{t\in[k]}\sum_{i\in C_t}\left\|x_i - \frac{1}{|C_t|}\sum_{j\in C_t} x_j\right\|^2 \quad \text{subject to} \quad C_1 \sqcup \cdots \sqcup C_k = T.$$

While this optimization problem is NP-hard to solve (even in $m = 2$ dimensions [6]), real-world instances of this problem are frequently solved using Lloyd's algorithm, which alternates between computing centroids and reassigning points to the nearest centroid. A data scientist can perform Lloyd's algorithm with random initializations to produce locally optimal clusterings, but when should he stop looking for a better clustering?

One popular initialization for Lloyd's algorithm is k-means++, which randomly selects k initial "centroids" from $\{x_i\}_{i\in T}$ in a way that encourages different centroids to be far apart. Letting W denote the random value of the k-means++ initialization, then the main result of [1] gives that

$$\text{val}(T\text{-IP}) \geq \frac{1}{8(\log k + 2)} \cdot \mathbb{E}W.$$

While this gives an approximation guarantee for the optimal clustering, it appears to be quite loose. For example, running several trials of k-means++ on the MNIST training set of 60,000 handwritten digits [4] with $k = 10$ produces a clustering with value about 39.22, whereas estimating $\mathbb{E}W$ leads to a lower bound of about 2.15.

In pursuit of a better lower bound, one may consider the Peng–Wei SDP relaxation:

(T-SDP)

$$\text{minimize} \quad \frac{1}{2|T|}\operatorname{tr}(DX) \quad \text{subject to} \quad \operatorname{tr}(X) = k,\ X1 = 1,\ X \geq 0,\ X \succeq 0.$$

Here, D denotes the $T \times T$ matrix whose (i,j)th entry is $\|x_i - x_j\|^2$, whereas $X \geq 0$ ensures that X is entrywise nonnegative and $X \succeq 0$ ensures that X is symmetric and positive semidefinite. For any clustering $C_1 \sqcup \cdots \sqcup C_k = T$, the

matrix $X = \sum_{t\in[k]} \frac{1}{|C_t|} 1_{C_t} 1_{C_t}^\top$ is feasible in (T-SDP) with SDP value equal to its IP value, and so val(T-SDP) is a lower bound for val(T-IP), as desired. Moreover, there has been a lot of work recently to establish how good this lower bound is for various random data models [2, 3, 8, 5]. However, SDPs are notoriously slow for large problem instances, and so it is infeasible to compute this lower bound for the MNIST training set (say).

To decrease the complexity of the problem, we can pass to a subset of the data. For example, pick $s \le |T|$ and then draw S uniformly from all subsets of T of size s. Then letting $C_1^* \sqcup \cdots \sqcup C_k^* = T$ denote the (T-IP)-optimal clustering, we have

$$\begin{aligned}\mathbb{E}\,\mathrm{val}(S\text{-SDP}) \le \mathbb{E}\,\mathrm{val}(S\text{-IP}) &\le \mathbb{E}\left[\frac{1}{s}\sum_{t\in[k]}\sum_{i\in C_t^*\cap S}\left\|x_i - \frac{1}{|C_t^*\cap S|}\sum_{j\in C_t^*\cap S} x_j\right\|^2\right]\\ &\le \mathbb{E}\left[\frac{1}{s}\sum_{t\in[k]}\sum_{i\in C_t^*\cap S}\left\|x_i - \frac{1}{|C_t^*|}\sum_{j\in C_t^*} x_j\right\|^2\right]\\ &= \mathrm{val}(T\text{-IP}).\end{aligned}$$

As such, we can produce a lower bound on val(T-IP) by estimating $\mathbb{E}\,\mathrm{val}(S$-SDP), which is computationally feasible when s is small. For example, if we select $s = 200$, then $\mathbb{E}\,\mathrm{val}(S\text{-SDP}) \approx 35$ for the MNIST training set, meaning the clustering from k-means++ is within 15 percent of optimal. In [7], we prove that this approach leads to a 99%-confidence 3-approximation certificate for any mixture of two spherical Guassians, and furthermore, this certificate can be computed in sub-linear time.

For follow-on work, it would be helpful to understand the distribution of val(S-SDP), as this would allow us to use even better test statistics for our certificate. Also, how does our approximation ratio scale with k? Can we extrapolate a near-optimal clustering for T from the SDP of S? Do our techniques transfer to other SDPs to provide sub-linear bounds? We are very interested to pursue these directions further.

References

[1] D. Arthur, S. Vassilvitskii, k-means++: The advantages of careful seeding, SODA (2017) 1027–1035.

[2] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, R. Ward, Relax, no need to round: Integrality of clustering formulations, ITCS (2015) 191–200.

[3] T. Iguchi, D. G. Mixon, J. Peterson, S. Villar, Probably certifiably correct k-means clustering, Math. Program. 165 (2017) 605–642.

[4] Y. LeCun, C. Cortes, C. J. C. Burges, The MNIST database of handwritten digits, `http://yann.lecun.com/exdb/mnist/`

[5] X. Li, Y. Li, S. Ling, T. Strohmer, K. Wei, When Do Birds of a Feather Flock Together? K-Means, Proximity, and Conic Programming, arXiv:1710.06008

[6] M. Mahajan, P. Nimbhorkar, K. Varadarajan, The planar k-means problem is NP-hard, Theor. Comput. Sci. 442 (2012) 13–21.
[7] D. G. Mixon, S. Villar, Monte Carlo approximation certificates for k-means clustering, arXiv:1710.00956
[8] D. G. Mixon, S. Villar, R. Ward, Clustering subgaussian mixtures by semidefinite programming, Inform. Inferenc 6 (2017) 389–415.
[9] J. Peng, Y. Wei, Approximating k-means-type clustering via semidefinite programming, SIAM J. Optimiz. 18 (2007), 186–205.

Reporter: Chen-Yun Lin

Participants

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

Prof. Dr. Radu V. Balan
Department of Mathematics
University of Maryland
College Park, MD 20742-4015
UNITED STATES

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. Jean-Luc Bouchot
Lehrstuhl für Mathematik C (Analysis)
RWTH Aachen
Pontdriesch 10
52062 Aachen
GERMANY

Dr. Claire Boyer
Laboratoire de Probabilités, Statistique et Modélisation (LPSM)
Sorbonne Université
Case 247
4, Place Jussieu
75252 Paris Cedex 05
FRANCE

Prof. Dr. A. Robert Calderbank
Pratt School of Engineering
Duke University
Durham, NC 27708
UNITED STATES

Dr. Maria Charina
Institut für Mathematik
Universität Wien
Oskar-Morgenstern-Platz 1
1090 Wien
AUSTRIA

Prof. Dr. Yuxin Chen
Department of Electrical Engineering
Princeton University
Engineering Quadrangle
Olden Street
Princeton, NJ 08544
UNITED STATES

Prof. Dr. Alexander Cloninger
Department of Mathematics
University of California, San Diego
9500 Gilman Road
San Diego, CA 92122
UNITED STATES

Prof. Dr. Albert Cohen
Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie
4, Place Jussieu
75005 Paris Cedex
FRANCE

Dr. Nadav Cohen
School of Mathematics
Institute for Advanced Study
1, Einstein Drive
Princeton, NJ 08540
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
Lehrstuhl für Mathematik C (Analysis)
RWTH Aachen
Templergraben 55
52062 Aachen
GERMANY

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

Prof. Dr. Simon Foucart
Department of Mathematics
Texas A & M University
College Station, TX 77843-3368
UNITED STATES

Prof. Dr. Tingran Gao
Department of Statistics
University of Chicago
Jones 316
5747 S. Ellis Avenue
Chicago IL 60637-1441
UNITED STATES

Martin Genzel
Institut für Mathematik
Sekr. MA 5-4
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin
GERMANY

Prof. Dr. Remi Gribonval
INRIA Rennes
Campus Beaulieu
35042 Rennes Cedex
FRANCE

Prof. Dr. Karlheinz Gröchenig
Fakultät für Mathematik
Universität Wien
Oskar-Morgenstern-Platz 1
1090 Wien
AUSTRIA

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
50937 Köln
GERMANY

Prof. Dr. Babak Hassibi
Department of Electrical Engineering
California Institute of Technology
MC 136-93
1200 California Boulevard
Pasadena CA 91125
UNITED STATES

Dr. Dominik Jüstel
Institute for Biological and Medical Imaging
Helmholtz-Zentrum München
Ingolstädter Landstrasse 1
85764 Neuherberg
GERMANY

Sandra Keiper
Institut für Mathematik
Sekr. MA 5-4
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin
GERMANY

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

Christian Kümmerle
Zentrum Mathematik
Technische Universität München
Boltzmannstrasse 3
85748 Garching bei München
GERMANY

Dr. Richard Küng
Department of Theoretical Physics
California Institute of Technology
MC 105-50
1200 E. California Boulevard
Pasadena, CA 91125
UNITED STATES

Dr. Chen-Yun Lin
Department of Mathematics
Duke University
P.O.Box 90320
Durham, NC 27708-0320
UNITED STATES

Shuyang Ling
Courant Institute of Mathematical Sciences
New York University
251, Mercer Street
New York, NY 10012-1110
UNITED STATES

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

Nicholas Marshall
Department of Mathematics
Yale University
P.O. Box 208285
New Haven, CT 06520-8285
UNITED STATES

Maximilian März
Institut für Mathematik
Sekr. MA 5-4
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin
GERMANY

Prof. Dr. Hrushikesh N. Mhaskar
Institute for Mathematical Sciences
Claremont Graduate University
150 E. 10th Street
Claremont, CA 91125
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

Dr. Philipp Christian Petersen
Institut für Mathematik
Technische Universität Berlin
Sekr. MA 5-4
Straße des 17. Juni 136
10623 Berlin
GERMANY

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

Prof. Dr. Holger Rauhut
Lehrstuhl für Mathematik C (Analysis)
RWTH Aachen
Pontdriesch 10
52062 Aachen
GERMANY

Prof. Dr. Justin Romberg
School of Electrical and Computer Engineering
Georgia Institute of Technology
777, Atlantic Drive NW
Atlanta, GA 30332-0250
UNITED STATES

Prof. Dr. Oliver Schaudt
Lehrstuhl für Mathematik C (Analysis)
RWTH Aachen
Pontdriesch 10
52062 Aachen
GERMANY

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

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

Prof. Dr. Gabriele Steidl
Fachbereich Mathematik
Technische Universität Kaiserslautern
67653 Kaiserslautern
GERMANY

Alexander Stollenwerk
Lehrstuhl für Mathematik C (Analysis)
RWTH Aachen
Pontdriesch 10
52062 Aachen
GERMANY

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

Dr. Ronen Talmon
Department of Electrical Engineering
TECHNION -
Israel Institute of Technology
Haifa 32000
ISRAEL

Dr. Soledad Villar
NYU Center for Data Science
Office 621
60 5th Avenue
New York, NY 10011
UNITED STATES

Dr. Felix Voigtlaender
Mathematisch-Geographische Fakultät
Katholische Universität Eichstätt
Ostenstrasse 26-28
85072 Eichstätt
GERMANY