Applied Harmonic Analysis, Massive Data Sets, Machine Learning, and Signal Processing

Emmanuel Candès (Stanford University),
Ronald Coifman (Yale University),
Amit Singer (Princeton University),
Thomas Strohmer (University of California, Davis)

October 16 - 21, 2016

1 Overview of the Field

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.

The analysis of massive, high-dimensional, noisy, time-varying data sets has become a critical issue for a large number of scientists and engineers. 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 [4, 8, 6, 21, 1].

In particular the novel paradigm of sparsity and sparse approximations has had a tremendous impact on various areas in applied mathematics such as imaging sciences. It states that functions and signals which come from applications typically exhibit the property of admitting a representation in a suitable orthonormal basis or—more often due to the advantageous property of redundancy—a frame with very few non-zero coefficients or, in a weakened version, an approximation by such an expansion. Suitable representation systems were and are still being developed in the areas of applied harmonic analysis such as wavelets or curvelets.

Although compression schemes such as JPEG2000 might be considered the first breakthrough of this general approach, quite recently, the new area of compressive sensing revealed another applications with tremendous impact. Roughly speaking, it showed that signals exhibiting a sparse approximation can be recovered efficiently from what would previously have been considered highly incomplete measurements. This discovery has led to a canon of fundamentally new approaches for various previously considered almost insolvable problems, for instance, for signal and image recovery problems. Compressive sensing also paves

the way for data acquisition schemes that scale with the inherent information dimension instead of with the ambient dimension, thereby holding promise to mitigate the "Curse of Dimensionality".

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.

Graph Laplacians and related nonlinear mappings into low dimensional spaces have been shown to be powerful tools for organizing high dimensional data. Especially diffusion maps, which have their roots in harmonic analysis, have been a useful tool in reducing the dimensionality of the data as well as providing a measure for pattern recognition and feature extraction. They yield meaningful geometric descriptions of data sets for efficient representation of complex geometric structures. Diffusion wavelets merge the power of diffusion maps with the advantages of wavelets. They generalize classical wavelets, allowing for multiscale analysis and signal processing on general structures, such as manifolds, graphs and point clouds in Euclidean space. This has several applications, for instance in the study of data sets which can be modeled as graphs, and one is interested in learning functions on such graphs.

Another important development in this area is the Scattering Transform, developed by Stephane Mallat [16], which builds locally invariant, stable and informative signal representations by cascading wavelet modulus decompositions followed by a lowpass averaging filter. As we will discuss in more detail below, the scattering transform also establishes an important link between harmonic analysis and deep learning. Deep learning is making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. Yet, despite its success, so far we have very little theoretical understanding of what makes this approach work (or fail). Scattering transforms provide a promising line of attack for developing a theoretical framework for deep learning. By introducing the rich collection of tools from harmonic analysis into deep neural networks in a principled way, we should be able to greatly enhance the efficiency and performance of deep neural networks.

The aforementioned exciting new developments have not only further strengthened the connections between applied harmonic analysis, machine learning, and data mining, but they also set the stage for new disruptive ideas for analyzing and extracting knowledge from massive and complex data sets. The goal of this workshop was to ignite this new wave of developments. In the last decade we have witnessed significant advances in many individual core areas of data analysis, including machine learning, signal processing, statistics, optimization, and of course harmonic analysis. It appears highly likely that the next major breakthroughs will occur at the intersection of these disciplines. Hence, what is needed is a concerted effort to bring together world leading experts from all these areas, which was one of the aims of this workshop.

This workshop has revolved around the following topics:

(i) Emerging connections between harmonic analysis and deep learning;
(ii) Understanding the structure of high-dimensional data;
(iii) Construction of data-adaptive efficient representations;
(iv) Efficient algorithms for inverse problems on complex data sets.

2 Recent Developments and Open Problems

2.1 Emerging connections between deep learning and harmonic analysis

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 [12]. Deep neural networks build hierarchical invariant representations by applying a succession of linear and non-linear operators which are learned from training data. Deep neural networks, and in particular convolutional networks developed by LeCun [11, 13], have recently achieved state-of-the-art results on several complex object recognition tasks. In addition to beating records in image recognition, and speech recognition, it has beaten other machine-learning techniques at

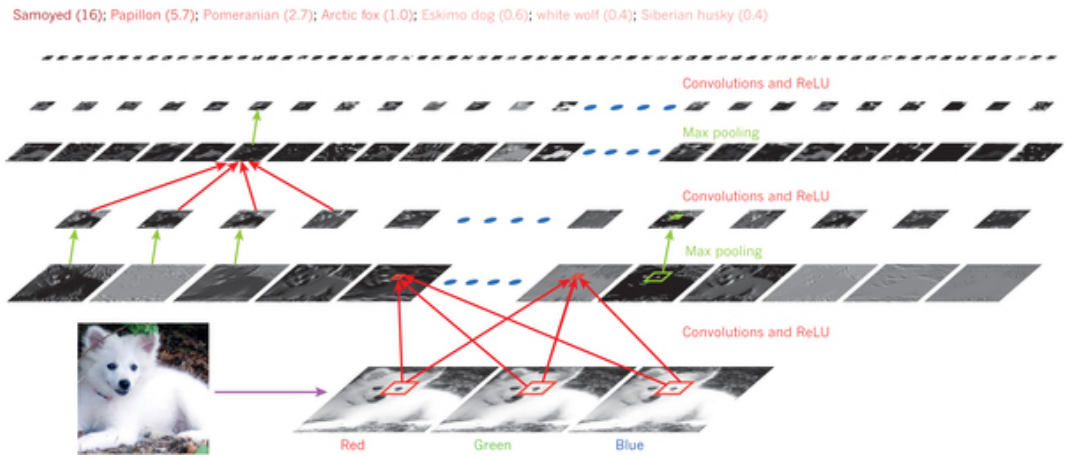

Figure 1: A convolutional network (from: Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature, vol. 521, no. 7553, pp. 436444, 2015.) The outputs of each layer of a typical convolutional network architecture applied to the image of a Samoyed dog (bottom left; and RGB (red, green, blue) inputs, bottom right). Each rectangular image is a feature map corresponding to the output for one of the learned features, detected at each of the image positions. Information flows bottom up, with lower-level features acting as oriented edge detectors, and a score is computed for each image class in the output.

predicting the activity of potential drug molecules, analyzing particle accelerator data, reconstructing brain circuits, and predicting the effects of mutations in non-coding DNA on gene expression and disease.

Convolutional nets are currently among the most successful deep learning architectures in a variety of tasks, in particular, in computer vision. A typical convolutional net used in computer vision applications consists of multiple convolutional layers, passing the input image through a set of filters followed by point-wise nonlinearity. An example architecture of a deep convolutional network is depicted in Figure 1.

A major issue in deep learning is to understand the properties of these networks, what needs to be learned and what is generic and common to most image classification problems. There are promising signs that this theoretical framework could be derived with tools from harmonic analysis. A first breakthrough towards this goal is the scattering transform, which has the structure of a convolutional network. Yet, rather than being learnt, the scattering network is obtained from the invariance, stability and informative requirements. A scattering transform builds invariant, stable and informative signal representations for classification. It is computed by scattering the signal information along multiple paths, with a cascade of wavelet modulus operators implemented in a deep convolutional network, see Figure 2. It is stable to deformations, which makes it particularly effective for image, audio and texture discrimination [3].

Scattering transforms provide a promising line of attack for developing a theoretical framework for deep learning. By introducing the rich collection of tools from harmonic analysis into deep neural networks in a principled way, we should be able to greatly enhance the efficiency and performance of deep neural networks.

While deep learning models have been particularly successful when dealing with signals such as speech, images, or video, in which there is an underlying Euclidean structure, recently there has been a growing interest in trying to apply learning on non-Euclidean geometric data, for example, in computer graphics and vision, natural language processing, and biology.

As Mallat points out in [17], supervised learning is a high-dimensional interpolation problem. We approximate a function $f(x)$ from q training samples $\{x_i, f(x_i)\}_{i=1}^q$, where x is a data vector of very high dimension d. In high dimension, x has a considerable number of parameters, which is a manifestation of the curse of dimensionality. Sampling uniformly a volume of dimension d requires a number of samples which grows exponentially with d. In most applications, the number q of training samples rather grows linearly with d. It is possible to approximate $f(x)$ with so few samples, only if f has some strong regularity properties allowing to ultimately reduce the dimension of the estimation. Any learning algorithm, including deep convolutional networks, thus relies on an underlying assumption of regularity. Specifying the nature of this

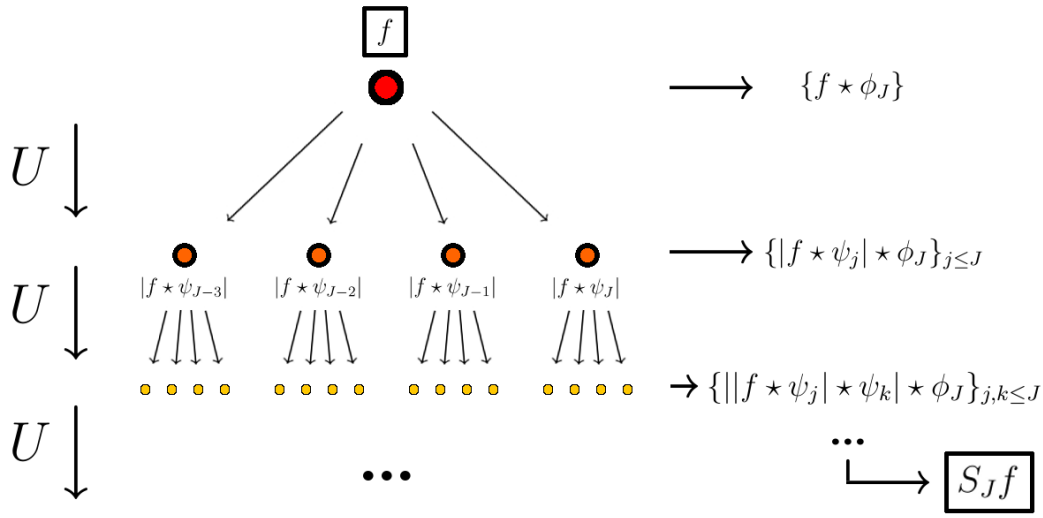

Figure 2: A scattering transform is computed by scattering the signal information along multiple paths, with a cascade of wavelet modulus operators implemented in a deep convolutional network. It is stable to deformations, which makes it particularly effective for image, audio and texture discrimination.

regularity is one of the core mathematical problems.

One can try to circumvent the curse of dimensionality by reducing the variability or the dimension of x, without sacrificing the ability to approximate $f(x)$. This is done by defining a new variable $\Psi(x)$ where Ψ is a contractive operator which reduces the range of variations of x, while still separating different values of f: $\Psi(x) \neq \Psi(x_0)$ if $f(x) \neq f(x_0)$. This separation-contraction trade-off needs to be adjusted to the properties of f. Linearization is a strategy used in machine learning to reduce the dimension with a linear projector. A low-dimensional linear projection of x can separate the values of f if this function remains constant in the direction of a high-dimensional linear space. This is rarely the case, but one can try to find $\Psi(x)$ which linearizes high-dimensional domains where $f(x)$ remains constant. The dimension is then reduced by applying a low-dimensional linear projector on $\Psi(x)$. Finding such a Ψ is the central goal of kernel learning algorithms.

Theory needs to be developed for Deep Learning to guide the search of proper feature extraction models at each layer. 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, and we do not know how to modify deep learning for those cases where it fails. How much training is really needed? And, perhaps one of the most difficult questions, how can we achieve unsupervised deep learning?

2.2 Understanding the structure of high-dimensional data

The need to analyze massive data sets in Euclidean space has led to a proliferation of research activity, including methods of dimension reduction and manifold learning. In general, understanding large data means identifying intrinsic characteristics of the data and developing techniques to isolate them.

While many of the currently existing tools (such as diffusion maps) show great promise, they rely on the assumption that data are stationary and homogeneous. Yet in many cases, we are dealing with changing and heterogeneous data. For instance, in medical diagnostics, we may want to infer a common phenomenon from data as diverse as MRI, EEG, and ECG. How do we properly fuse and process heterogeneous data to extract knowledge?

In a broad range of natural and real-world dynamical systems, measured signals are controlled by underlying processes or drivers. As a result, these signals exhibit highly redundant representations, while their temporal evolution can often be compactly described by dynamical processes on a low-dimensional manifold. Recently, diffusion maps have been generalized to the setting of a dynamic data set, in which the graph

associated with it changes depending on some set of parameters. The associated global diffusion distance allows measuring the evolution of the dynamic data set in its intrinsic geometry. However, this is just a first step. One objective of this workshop was dedicated to mathematical tools that can detect and capture in an automatic, unsupervised manner the inner architecture of large data sets.

2.3 Construction of data-adaptive efficient representations

Processing of signals on graphs is emerging as a fundamental problem in an increasing number of applications. Indeed, in addition to providing a direct representation of a variety of networks arising in practice, graphs serve as an overarching abstraction for many other types of data.

The construction of data-adaptive dictionaries is crucial, even more so in light of the need to analyze data that in past has not fallen within the boundary of signal processing, for example graphs or text documents. In fact, the above may be considered as casting a bridge between classical signal processing and the new era of processing of general data.

Convolutional neural networks have been successful in machine learning problems where the coordinates of the underlying data representation have a grid structure, and the data to be studied in those coordinates has translational equivariance/invariance with respect to this grid. However, e.g. data defined on 3-D meshes, such as surface tension or temperature, measurements from a network of meteorological stations, or data coming from social networks or collaborative filtering, are all examples of datasets on which one cannot apply standard convolutional networks. Clearly, this is another area where a closer link between deep learning, signal processing, and harmonic analysis would be highly beneficial.

2.4 Efficient algorithms for inverse problems on complex data sets

Inverse problems arising in connection with massive, complex data sets pose tremendous challenges and require new mathematical tools. Consider for instance femtosecond X-ray protein nanocrystallography. There the problem is to uncover the structure of (3-dimensional) proteins from multiple (2-dimensional) intensity measurements [23]. In addition to the huge amount of data and the fact that phase information gets lost during the measurement process, we also do not know the proteins' rotation, which change from illumination to illumination. Standard phase retrieval methods fail miserably in this case. Yet, recent advances at the intersection of harmonic analysis, optimization, and signal processing show promise to solve such challenging problems.

Other important inverse problems in this topic are tied to heterogenous data or to the idea of self-calibration. 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. Can we design efficient non-convex algorithms with provable convergence? One objective of this workshop was the advancement of new theoretical and numerical tools for such demanding inverse problems.

3 Presentation Highlights and Scientific Progress

In this section we discuss a few selected highlights among the many high caliber presentations. One of the key events of the workshop was the opening talk by Yann LeCun, who is famous for his pioneering work on Deep Learning and Artificial Intelligence. His presentation was aptly titled "Obstacles to AI, Mathematical and Otherwise". According to LeCun, prediction, perception, planning/reasoning, attention, and memory are the pillars of intelligence. Both animals and humans learn to predict, learn how the world works, and acquire common sense largely without supervision, through observation and experimentation. This is a far cry from supervised learning–the basis of most recent successes in the application of deep learning. Significant progress in AI will require breakthroughs in unsupervised/predictive learning, as well as in reasoning, attention, and episodic memory. Yann LeCun described several projects at FAIR and NYU on unsupervised learning for predicting videos using adversarial training, question answering with a new type of memory-augmented network, and various applications for vision and natural language understanding. In his talk,

Yann LeCu emphasized those open problems in artificial intelligence that are especially interesting for mathematicians. One of the problems is of course the general lack of a theoretical understanding behind why deep learning works and why it sometimes does not work. Despite the flurry of research activity in Deep Learning in recent years, almost no progress has been made regarding developing theory for Deep Learning.

Therefore, one focal point of this workshop was to close this gap between theory and practice of deep learning. Several presentations were devoted to this topic. Stephane Mallat, in his talk, was looking at the mathematical mysteries of deep networks. He presented an attempt to provide a partial answer to the aforementioned question about developing theory for deep learning [17]. Recall that multilayer neural networks are computational learning architectures which propagate the input data across a sequence of linear operators and simple non-linearities. The properties of shallow networks, with one hidden layer, are well understood as decompositions in families of ridge functions. However, these approaches do not extend to networks with more layers. Deep convolutional neural networks are implemented with linear convolutions followed by non-linearities, over typically more than 5 layers. These complex programmable machines, defined by potentially billions of filter weights, bring us to a different mathematical world. Many researchers have pointed out that deep convolution networks are computing progressively more powerful invariants as depth increases, but relations with networks weights and non-linearities are complex. Mallat aimed at clarifying important principles which govern the properties of such networks, but their architecture and weights may differ with applications. He showed that computations of invariants involve multiscale contractions, the linearization of hierarchical symmetries, and sparse separations. This conceptual basis is only a first, albeit a very important, step towards a full mathematical understanding of convolutional network properties.

Staying with this topic, Alexander Cloninger discussed the approximation of functions using deep neural nets. Given a function f on a d-dimensional manifold $\Gamma \in \mathbb{R}^m$, he constructed a sparsely-connected depth-4 neural network for which he was able to bound its error in approximating f, see [22]. The size of the network depends on dimension and curvature of the manifold Γ, the complexity of f, in terms of its wavelet description, and only weakly on the ambient dimension m. Essentially, the network computes wavelet functions, which are computed from so-called Rectified Linear Units. Hrushikesh Mhaskar analyzed the advantages of deep versus shallow networks, thereby settling an old conjecture by Bengio on the role of depth in networks. While the universal approximation property holds both for hierarchical and shallow networks, he proved that deep (hierarchical) networks can approximate the class of compositional functions with the same accuracy as shallow networks but with exponentially lower number of training parameters as well as VC-dimension [18]. He then introduced a general class of scalable, shift-invariant algorithms to show a simple and natural set of requirements that justify deep convolutional networks.

"The Deceptive Nature of Generalization in Machine Learning" was the topic of a lively presentation by Ben Recht. Preventing overfitting in machine learning is usually accomplished by constraining model size or adding regularizers to mitigate complexity. In his talk, Recht aimed to problematize these conventional techniques. He provided empirical evidence of function classes of high or infinite dimension that are able to achieve state-of-the-art performance on several benchmarks without any obvious forms of regularization. Such performance is possible even if the fit function exactly interpolates the training data. In response, borrowing tools from optimization and applied harmonic analysis, Recht attempted to establish a framework for generalization in machine learning that agrees with these experimental findings. This framework applies to standard high-dimensional linear models and strives to provide insights into contemporary neural network architectures.

Eero Simoncelli, in his talk "Cascaded gain control representations", analyzed a range of questions, including how populations of neurons extract and represent visual information and in what ways this is matched to, or optimized for, our visual environment, how do these representations enable or limit perception and what new principles may be gleaned from this representations and applied to engineered imaging or vision systems. He introduced a general framework for end-to-end optimization of the rate-distortion performance of nonlinear transform codes assuming scalar quantization. The framework can be used to optimize any differentiable pair of analysis and synthesis transforms in combination with any differentiable perceptual metric. As a concrete application he described an image compression system, consisting of a nonlinear encoding transformation, a uniform quantizer, and a nonlinear decoding transformation [2]. Like many deep neural network architectures, the transforms consist of layers of convolutional linear filters and nonlinear activation functions, but Simoncelli uses a joint nonlinearity that implements a form of local gain control, inspired by those used to model biological neurons. Using a variant of stochastic gradient descent, he jointly optimizes the system

for rate-distortion performance over a database of training images, introducing a continuous proxy for the discontinuous loss function arising from the quantizer. The relaxed optimization problem resembles that of variational autoencoders, except that it must operate at any point along the ratedistortion curve, whereas the optimization of generative models aims only to minimize entropy of the data under the model. The presented examples of compressed images were a striking demonstration of the power of the proposed framework.

While convex optimization has been a hot topic in the last decade, partly fueled by the field of compressive sensing and its extensions, very recently we have witnessed a strong trend towards developing a theoretical framework for algorithms for non-convex optimization problems, see e.g. [5]. These methods play a particularly important role in connection with massive datasets, since they promise improved numerical efficiency compared to their "convex cousins". This workshop provided a platform for recent developments in this important topic.

John Wright, in his talk about "Nonconvex Recovery of Low-Complexity Models", considered a complete dictionary recovery problem, in which we are given a data matrix Y, and the goal is to factor it into a product $Y \sim A_0 X_0$, where A_0 is a square and invertible matrix and X_0 is a sparse matrix of coefficients. This is an abstraction of the dictionary learning problem, in which we try to learn a concise approximation to a given dataset. While dictionary learning is widely used in signal processing and machine learning, relatively little is known about it in theory. Much of the difficulty owes to the fact that standard learning algorithms solve nonconvex problems, and are difficult to analyze globally. The talk described an efficient algorithm which provably learns representations in which the matrix X_0 has as many as $\mathcal{O}(n)$ nonzeros per column, under a suitable probability model for X_0. Previous efficient algorithms either only worked for very sparse instances or required multiple rounds of SDP relaxation. Wright's results follow from a reformulation of the dictionary recovery problem as a nonconvex optimization over a high dimensional sphere. This particular nonconvex problem has a surprising property: once about n^3 data samples have been observed, with high probability the objective function has no spurious local minima [24]. This geometric phenomenon, in which seemingly challenging nonconvex problems can be solved globally by efficient iterative methods, also arises in problems such as tensor decomposition and phase recovery from magnitude measurements. Wright sketched these connections and illustrated his results with applications in microscopy and computer vision.

Clustering is a central problem in unsupervised machine learning. It consists of partitioning a given finite sequence of points $\{x_i\}_{i=1}^N$ in $\mathbb{R}^m$ into k subsequences such that some dissimilarity function is minimized. Usually, this function is chosen ad hoc with an application in mind. A particularly common choice is the k-means objective. A popular heuristic for solving k-means is Lloyd's algorithm, also known as the k-means algorithm. This algorithm alternates between calculating centroids of proto-clusters and reassigning points according to the nearest centroid. In general, Lloyd's algorithm (and its variants) may converge to local minima of the k-means objective. Furthermore, the output of Lloyd's algorithm does not indicate how far it is from optimal. Thus, there is an urgent need for rigorous mathematical theory to put k-means and its variations on a more solid footing. In her talk, Soledad Villar presented some significant recent progress towards this goal [19]. She introduced a model-free relax-and-round algorithm for k-means clustering based on a semidefinite relaxation due to Peng and Wei [20]. The algorithm interprets the SDP output as a denoised version of the original data and then rounds this output to a hard clustering. Her approach provides a generic method for proving performance guarantees for this algorithm, and allows one to analyze the algorithm in the context of subgaussian mixture models. Villar also studied the fundamental limits of estimating Gaussian centers by k-means clustering in order to compare her approximation guarantee to the theoretically optimal k-means clustering solution.

Blind deconvolution is another important problem, that arises in many applications but still lacks a rigorous framework for the analysis of efficient algorithms. More precisely, in blind deconvolution one studies the question of reconstructing two signals f and g from their convolution $y = f * g$. This problem, known as blind deconvolution, pervades many areas of science and technology, including astronomy, medical imaging, optics, and wireless communications. A key challenge of this intricate non-convex optimization problem is that it might exhibit many local minima. Shuyang Ling, in his talk "Rapid, Robust, and Reliable Blind Deconvolution via Nonconvex Optimization" presented an efficient numerical algorithm that is guaranteed to recover the exact solution, when the number of measurements is (up to log-factors) slightly larger than the information-theoretical minimum, and under reasonable conditions on f and g. The proposed regularized gradient descent algorithm converges at a geometric rate and is provably robust in the presence of noise [15]. His algorithm is arguably the first blind deconvolution algorithm that is numerically efficient, robust against

noise, and comes with rigorous recovery guarantees under certain subspace conditions. Moreover, numerical experiments do not only provide empirical verification of his theory, but they also demonstrate that our method yields excellent performance even in situations beyond the theoretical framework.

Various applications involve assigning discrete label values to a collection of objects based on some noisy data. Due to the discreteand hence nonconvexstructure of the problem, computing the maximum likelihood estimates becomes intractable at first sight. Yuxin Chen presented recent progress towards efficient computation of the maximum likelihood estimates by focusing on a concrete joint alignment problemthat is, the problem of recovering n discrete variables $x_i \in \{1, \dots, m\}, 1 \leq i \leq n$, given noisy observations of their modulo differences $\{x_i - x_j \mod m\}$. He proposed a novel low-complexity procedure, which operates in a lifted space by representing distinct label values in orthogonal directions, and which attempts to optimize quadratic functions over hyper cubes. Starting with a first guess computed via a spectral method, the algorithm successively refines the iterates via projected power iterations. Chen proved that the proposed projected power method converges to the maximum likelihood estimate in a suitable regime [7]. The practicality of the proposed algorithm was illustrated via numerical experiments for both synthetic and real data.

Fast numerical algorithms are a major concern when dealing with massive data sets. This topic was addresses in several talks such as the aforementioned presentations on efficient algorithms for non-convex optimization, as well as in a presentation by Alexandre d'Aspremont. He described a powerful regularized nonlinear acceleration technique for generic optimization problems. The proposed scheme computes estimates of the optimum from a nonlinear average of the iterates produced by any optimization method. The weights in this average are computed via a simple linear system, whose solution can be updated online. This acceleration scheme runs in parallel to the base algorithm, providing improved estimates of the solution on the fly, while the original optimization method is running. Numerical experiments, detailed on classical classification problems, demonstrated the efficacy of the proposed framework.

Functional maps and functional map networks for joint data analysis was the topic of the talk by Leonid Guibas. The construction of networks of maps among shapes in a collection enables a variety of applications in data-driven geometry processing. A key task in network construction is to make the maps consistent with each other. This consistency constraint, when properly defined, leads not only to a concise representation of such networks, but more importantly, it serves as a strong regularizer for correcting and improving noisy initial maps computed between pairs of shapes in isolation. Up-to-now, however, the consistency constraint has only been fully formulated for point-based maps or for shape collections that are fully similar. Guibas, in his talk, introduced a framework for computing consistent functional maps within heterogeneous shape collections [10]. In such collections not all shapes share the same structuredifferent types of shared structure may be present within different (but possibly overlapping) sub-collections. Unlike point-based maps, functional maps can encode similarities at multiple levels of detail (points or parts), and thus are particularly suitable for coping with such diversity within a shape collection. He showed how to rigorously formulate the consistency constraint in the functional map setting. The formulation leads to a powerful tool for computing consistent functional maps, and also for discovering shared structures, such as meaningful shape parts.

One of the challenges in data analysis is to distinguish between different sources of variability manifested in data. In his presentation, Ronen Talmon considered the case of multiple sensors measuring the same physical phenomenon, such that the properties of the physical phenomenon are manifested as a hidden common source of variability (which we would like to extract), while each sensor has its own sensor-specific effects. He presented a method based on alternating products of diffusion operators, and showed that it extracts the common source of variability [14]. Moreover, this method extracts the common source of variability in a multi-sensor experiment as if it were a standard manifold learning algorithm used to analyze a simple single-sensor experiment, in which the common source of variability is the only source of variability.

Another challenge in data analysis is to robustly model both high-dimensional and massive datasets by a low-rank subspace. This problem was attacked by Gilad Lerman in his talk. Assume we are given high-dimensional data (large p, large n regime) sampled from a generalized elliptical distribution (possibly with heavy tails) and establish effective recovery of the shape of its covariance in that regime. The shape of the covariance can then be used for robust subspace modeling. Moreover, we may consider a massive data set in an ad hoc wireless sensor network, where each node has access to only one chunk of the dataset. We further assume no central processing (e.g. due to physical or security restrictions). Lerman proposed a distributed solution for robust subspace modeling with a fast convergence rate and recovery guarantees (when there is a generative model). The main mathematical contribution is the careful solution of a local dual minimization

problem. He discussed various open problems of both scenarios and their extensions, obstacles in unifying them and obstacles in obtaining numerically efficient and theoretically guaranteed algorithms in very high dimensions.

Feature selection from real-world data with nonlinear observations was the focus of a presentation by Gitta Kutyniok [9]. Assume we need to select features n based on a relatively small collection of sample pairs $(x_i, y_i)_{i=1,\dots,m}$. The observations $y_i \in \mathbb{R}$ are supposed to follow a noisy single-index model, depending on a certain set of signal variables. A major difficulty is that these variables usually cannot be observed directly, but rather arise as hidden factors in the actual data vectors $x_i \in \mathbb{R}^d$ (feature variables). Kutyniok showed that a successful variable selection is still possible in this setup, even when the applied estimator does not have any knowledge of the underlying model parameters and only takes the raw samples (x_i, y_i) as input. The model assumptions were fairly general, allowing for non-linear observations, arbitrary convex signal structures as well as strictly convex loss functions. This is particularly appealing for practical purposes, since in many applications, already standard methods, e.g., the Lasso or logistic regression, yield surprisingly good outcomes. The versatility of this framework was impressively demonstrated by means of a specific real-world problem, namely sparse feature extraction from (proteomics-based) mass spectrometry data.

4 Outcome of the meeting

Based on the quality of presentations, the intense scientific collaborations, and the enthusiastic feedback from the participants, this workshop was hugely successful in bringing together world leading experts at the intersection of applied harmonic analysis, large data sets, machine learning, and signal processing to present recent developments, in fostering new cooperations, and in making significant progress, or at least paving the way, towards solving some of the problems described in the previous sections. At the same time, the passionate discussions and focused interactions during this workshop have perhaps produced as many questions as they produced answers. On the other hand, articulating meaningful and precise questions is often the most important step towards scientific breakthroughs.

References

[1] William K Allard, Guangliang Chen, and Mauro Maggioni. Multi-scale geometric methods for data sets ii: Geometric multi-resolution analysis. *Applied and Computational Harmonic Analysis*, 32(3):435–462, 2012.

[2] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

[3] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[4] Emmanuel J. Candès and David L. Donoho. Curvelets and curvilinear integrals. *J. Approx. Theory*, 113(1):59–90, 2001.

[5] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007, 2015.

[6] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[7] Yuxin Chen and Emmanuel Candes. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *arXiv preprint arXiv:1609.05820*, 2016.

[8] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc.Natl.Acad.Sci. USA*, 102(21):7426–7431, 2005.

[9] Martin Genzel and Gitta Kutyniok. A mathematical framework for feature selection from real-world data with non-linear observations. *arXiv preprint arXiv:1608.08852*, 2016.

[10] Qixing Huang, Fan Wang, and Leonidas Guibas. Functional map networks for analyzing and exploring large shape collections. *ACM Transactions on Graphics (TOG)*, 33(4):36, 2014.

[11] Y. Le Cun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W Hubbard, and L.D. Jackel. Handwritten digit recognition with a back-propagation network. In *Proc. Advances in Neural Information Processing Systems*, pages 394–404, 1990.

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] Roy R Lederman and Ronen Talmon. Common manifold learning using alternating-diffusion. Technical report, submitted, Tech. Report YALEU/DCS/TR1497, 2014.

[15] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv preprint arXiv:1606.04933*, 2016.

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

[17] Stéphane Mallat. Understanding deep convolutional networks. *Phil. Trans. R. Soc. A*, 374(2065):2015.0203, 2016.

[18] Hrushikesh N Mhaskar and Tomaso Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.

[19] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *arXiv preprint arXiv:1602.06612*, 2016.

[20] Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205, 2007.

[21] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization. *SIAM Review*, 52: 471–501, 2010.

[22] Uri Shaham, Alexander Cloninger, and Ronald R Coifman. Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis*, 2016.

[23] Amit Singer, Zhizhen Zhao, Yoel Shkolnisky, and Ronny Hadani. Viewing angle classification of cryo-electron microscopy images using eigenvectors. *SIAM Journal on Imaging Sciences*, 4(2):723–759, 2011.

[24] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I: Overview and geometric picture. *preprint*, http://arxiv.org/abs/1504.06785, 2015.