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 27- November 1, 2019

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.

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. Moreover, developments in applied harmonic analysis often have ignited research in other areas, such as the recent surge in rigorous studies of non-convex optimization methods.

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.

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). By introducing the rich collection of tools from harmonic analysis into deep neural networks in a principled way, we should be able to gain some theoretical insight into the deep learning "black box", and thereby 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) Connections between harmonic analysis and deep learning;

(ii) Understanding the structure of high-dimensional and multimodal data and the construction of data-adaptive efficient representations;

(iii) 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 decade 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 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 pointwise nonlinearity.

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. It is stable to deformations, which makes it particularly effective for image, audio and texture discrimination [3].

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 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.

Data are inherently heterogeneous, as they may come in different modalities, such as text, sound, pictures, and geolocation. Yet this data needs to be processed and analyzed in an integrated manner to make optimal use of the available information. Unlike traditional unimodal learning systems, multimodal systems can carry complementary information about each other, which will only become evident when all modalities are included in the learning process. While multimodal learning is a rather difficult topic, the task of generating multimodal synthetic data poses even more challenges. Current methods in machine learning are often not suitable to address multimodal data. Developing algorithms that can handle multimodal data was another important topic in this workshop.

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, low-rankness, 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 Marco Cuturi on *Computational Optimal Transport*. Optimal transport provides a powerful and flexible way to compare probability measures, of all shapes: absolutely continuous, degenerate, or discrete. This includes of course point clouds, histograms of features, and more generally datasets, parametric densities or generative models. Originally proposed by Monge in the eighteenth century, this theory later led to Nobel Prizes for Koopmans and Kantorovich as well as Villani's and Figalli's Fields Medals in 2010 and 2018. After having attracted the interest of mathematicians for several years, optimal transport has recently reached the machine learning community, because it can now tackle (both in theory and numerically) challenging learning scenarios, including for instance dimensionality reduction and structured prediction problems that involve histograms or point clouds, and estimation of parametric densities or generative models in highly degenerate / high-dimensional problems. In his talk, Cutiru gave a short introduction to optimal transport theory and explained how these intuitive tools require some adaptation before being used in high-dimensional settings. He presented computational strategies to cope with the challenges of scale arising from the use of this theory on real-life problems.

An important focus of the workshop was emerging theory for deep learning. Stephane Mallat, in his talk, was looking at the mathematical mysteries of deep networks. A major issue in harmonic analysis is to capture the phase dependence of frequency representations, which carries important signal properties. It seems that

convolutional neural networks have found away. Over time-series and images, convolutional networks often learn a first layer of filters which are well localized in the frequency domain, with different phases. Mallat showws that a rectifier then acts as a filter on the phase of the resulting coefficients. It computes signal descriptors which are local in space, frequency and phase. The non-linear phase filter becomes a multiplicative operator over phase harmonics computed with a Fourier transform along the phase. He proved that it defines a bi-Lipschitz and invertible representation. The correlations of phase harmonics coefficients characterise coherent structures from their phase dependence across frequencies. Several impressive numerical simulations complemented the talk.

Gitta Kutyniok's presentation was concerned with *The Approximation Power of Deep Neural Networks: Theory and Applications*. As mentioned before, despite the outstanding success of deep neural networks in real-world applications, most of the related research is empirically driven and a mathematical foundation is almost completely missing. The main goal of a neural network is to approximate a function, which for instance encodes a classification task. Thus, one theoretical approach to derive a fundamental understanding of deep neural networks focusses on their approximation abilities. In her talk Kutyniok discussed theoretical results which prove that not only do (memory-optimal) neural networks have as much approximation power as classical systems such as wavelets or shearlets, but they are also able to beat the curse of dimensionality. On the numerical side, she will showed that superior performance can typically be achieved by combining deep neural networks with classical approaches from approximation theory.

Staying with this theme, Helmut Bölcskei talked about *Fundamental limits of deep neural network learning*. He developed the fundamental limits of learning in deep neural networks by characterizing what is possible if no constraints on the learning algorithm and the amount of training data are imposed. Concretely, he considered Kolmogorov-optimal approximation through deep neural networks with the guiding theme being a relation between the complexity of the function (class) to be approximated and the complexity of the approximating network in terms of connectivity and memory requirements for storing the network topology and the associated quantized weights. The developed theory educes remarkable universality properties of deep networks. Specifically, deep networks are optimal approximants for vastly different function classes such as affine systems and Gabor systems. In addition, deep networks provide exponential approximation accuracy of widely different functions including the multiplication operation, polynomials, sinusoidal functions, general smooth functions, and even one-dimensional oscillatory textures and fractal functions such as the Weierstrass function, both of which do not have any known methods achieving exponential approximation accuracy.

Mauro Maggioni, in his talk *Learning Interaction laws in particle- and agent-based systems* focused on the following inference problem for a system of interacting particles or agents: given only observed trajectories of the system, we are interested in estimating the interaction laws between the particles/agents. Hr considered both the mean-field limit (i.e. the number of particles going to infinity) and the case of a finite number of agents, with an increasing number of observations. It was shown that at least in the particular setting where the interaction is governed by an (unknown) function of pairwise distances, under a suitable coercivity condition that guarantees the well-posedness of the problem of recovering the interaction kernel, statistically and computationally efficient, nonparametric, suitably-regularized least-squares estimators exist. Furthermore, the audience learned that the high-dimensionality of the state space of the system does not affect the learning rates, and our estimators achieve the optimal learning rate for one-dimensional (the variable being pairwise distance) regression problems with noisy observations. Efficient algorithms for constructing the estimator for the interaction kernels were presented, with statistical guarantees, and demonstrate them on various simple examples, including extensions to agent systems with different types of agents, second-order systems, and families of systems with parametric interaction kernels.

Tingran Gao gave an compelling presentation on *Multi-Representation Manifold Learning on Fibre Bundles*. Spectral geometry has played an important role in modern geometric data analysis, where the technique is widely known as Laplacian eigenmaps or diffusion maps. In his talk, he presented a geometric framework that studies graph representations of complex datasets, where each edge of the graph is equipped with a non-scalar transformation or correspondence. This new framework models such a dataset as a fibre bundle with a connection, and interprets the collection of pairwise functional relations as defining a horizontal diffusion process on the bundle driven by its projection on the base. The eigenstates of this horizontal diffusion process encode the "consistency" among objects in the dataset, and provide a lens through which the geometry of the dataset can be revealed. Gao demonstrated an application of this geometric framework on evolutionary

anthropology.

Data visualization is a long-standing topic in data analysis, which has witnessed some exciting progress in recent years. Guy Wolf, in his talk *Geometry-based Data Exploration* presented some of these developments. High-throughput data collection technologies are becoming increasingly common in many fields, especially in biomedical applications involving single cell data (e.g., scRNA-seq and CyTOF). These introduce a rising need for exploratory analysis to reveal and understand hidden structure in the collected (high-dimensional) Big Data. A crucial aspect in such analysis is the separation of intrinsic data geometry from data distribution, as (a) the latter is typically biased by collection artifacts and data availability, and (b) rare subpopulations and sparse transitions between meta-stable states are often of great interest in biomedical data analysis. Wolf showed several tools that leverage manifold learning, graph signal processing, and harmonic analysis for biomedical (in particular, genomic/proteomic) data exploration, with emphasis on visualization, data generation/augmentation, and nonlinear feature extraction. A common thread in the presented tools is the construction of a data-driven diffusion geometry that both captures intrinsic structure in data and provides a generalization of Fourier harmonics on it. These, in turn, are used to process data features along the data geometry for denoising and generative purposes. He related this approach to the recently-proposed geometric scattering transform that generalizes Mallat's scattering to non-Euclidean domains, and provides a mathematical framework for theoretical understanding of the emerging field of geometric deep learning.

Stefan Steinerberger talked about *"Dimensionality Reduction via tSNE – the mathematical theory and the remaining challenges"*. tSNE has become the standard method of dimensionality reduction in the medical sciences (clustering cell types, gene expressions, etc.); amusingly (or maybe not amusingly), the mathematical theory did not exist until recently. Even a basic understanding of the mathematics immediately implies a series of improvements. In his very entertaining talk, Steinerberger surveyed the existing arguments and discussed a number of interesting problems; given the prevalence of tSNE in medicine, even small improvements can affect a lot of research within a very short period of time.

Data clustering is another workhorse in machine learning and data science. Dustin Mixon, in *"SqueezeFit: Label-aware dimensionality reduction by semidefinite programming"* presented exciting progress in this area. Given labeled points in a high-dimensional vector space, we seek a low-dimensional subspace such that projecting onto this subspace maintains some prescribed distance between points of differing labels. Intended applications include compressive classification. His talk introduced a semidefinite relaxation of this problem, along with various performance guarantees.

Shuyang Ling considered related problems in *"Group synchronization on complex networks: nonconvex optimization and Kuramoto model"*. Information retrieval from graphs plays an increasingly important role in data science and machine learning. His talk focused on two such examples. The first one concerns the graph cuts problem: how to find the optimal k-way graph cuts given an adjacency matrix. Ling presented a convex relaxation of ratio cut and normalized cut, which gives rise to a rigorous theoretical analysis of graph cuts. He derived remarkable deterministic bounds of finding the optimal graph cuts via a spectral proximity condition which naturally depends on the intra-cluster and inter-cluster connectivity. Moreover, his theory even provides theoretic guarantees for spectral clustering and community detection under stochastic block model. The second example is about the landscape of a nonconvex cost function arising from group synchronization and matrix completion. This function also appears as the energy function of coupled oscillators on networks. We study how the landscape of this function is related to the underlying network topologies. Ling proved that the optimization landscape has no spurious local minima if the underlying network is a deterministic dense graph or an Erdos-Renyi random graph. The results find applications in signal processing and dynamical systems on networks.

Several exciting applications were at the center of several outstanding presentations. For example, Gal Mishne talked about *Multiway tensor analysis in neuroscience*. Experimental advances in neuroscience enable the acquisition of increasingly large-scale, high-dimensional and high-resolution neuronal and behavioral datasets, however addressing the full spatiotemporal complexity of these datasets poses significant challenges for data analysis and modeling. She proposed to model such datasets as multiway tensors with an underlying graph structure along each mode, learned from the data. Mishne presented three frameworks we have developed to model, analyze and organize tensor data that infer the coupled multi-scale structure of the data, reveal latent variables and visualize short and long-term temporal dynamics with applications in calcium imaging analysis, fMRI and artificial neural networks

Roy Lederman spoke about *On the Continuum Between Models, Data-Driven Discovery and Machine*

Learning: Mapping the Continuum of Molecular Conformations Using Cryo-Electron Microscopy. Cryo-Electron Microscopy (cryo-EM) is an imaging technology that is revolutionizing structural biology. Cryo-electron microscopes produce a large number of very noisy two-dimensional projection images of individual frozen molecules; unlike related methods, such as computed tomography (CT), the viewing direction of each particle image is unknown. The unknown directions, together with extreme levels of noise and additional technical factors, make the determination of the structure of molecules challenging. While other methods for structure determination, such as x-ray crystallography and nuclear magnetic resonance (NMR), measure ensembles of molecules, cryo-electron microscopes produce images of individual molecules. Therefore, cryo-EM could potentially be used to study mixtures of different conformations of molecules. Indeed, current algorithms have been very successful at analyzing homogeneous samples, and can recover some distinct conformations mixed in solutions, but, the determination of multiple conformations, and in particular, continuums of similar conformations (continuous heterogeneity), remains one of the open problems in cryo-EM. In practice, some of the key components in "molecular machines" are flexible and therefore appear as very blurry regions in 3-D reconstructions of macro-molecular structures that are otherwise stunning in resolution and detail. Ledermann discussed "hyper-molecules," the mathematical formulation of heterogenous 3-D objects as higher dimensional objects, and the machinery that goes into recovering these "hyper-objects" from data. He analyzed some of the statistical and computational challenges, and how they are addressed by merging data-driven exploration, models and computational tools originally built for deep-learning.

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.