Computational Harmonic Analysis in Data Science and Machine Learning

Xiuyuan Cheng (Duke University),
Amit Singer (Princeton University),
Thomas Strohmer (University of California, Davis),
Soledad Villar (John Hokpins University)

September 15 - 20, 2024

1 Overview of the Field

Future progress in science and technology depends crucially on our ability to derive new discoveries and deep understanding from complex and massive datasets. However, uncertainty, scale, non-stationarity, noise, and heterogeneity are fundamental issues impeding progress at all phases of the pipeline that creates knowledge from data. New mathematical challenges arise as current algorithms are in many cases no longer able to keep up with the numerous demands and changing environments as well as the huge amounts of data that need to be processed and analyzed. Fundamentally new ideas and approaches are needed and will require a close collaboration between mathematicians, statisticians, computer scientists, and engineers.

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. Compressive sensing [2, 1, 3], low-rank matrix completion, provable methods for phase retrieval, multiscale graph signal processing, and the scattering transform [6] are just a few developments of applied harmonic analysis that has significantly impacted the fields of inverse problems, image processing, machine learning, and data science in recent years [17, 16].

The aforementioned exciting new developments have not only further strengthened the connections between applied harmonic analysis, machine learning, and AI, 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 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.

(iv) Randomized algorithms in data science and machine learning
(v) Manifold learning, Wasserstein space, and related topics

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 or so is the emergence of *deep learning* [4]. Deep learning has had a transformative impact across a wide range of applications, revolutionizing industries and creating new possibilities for innovation. One of the most notable areas where deep learning has made a significant impact is computer vision. Deep learning models, particularly convolutional neural networks (CNNs), have dramatically improved image recognition, object detection, and segmentation tasks. These advancements have enabled applications like facial recognition, medical imaging analysis, autonomous vehicles, and even augmented reality. In medicine, for instance, deep learning is being used to detect diseases in radiology images, such as tumors or abnormalities in X-rays, CT scans, and MRIs, often with accuracy that rivals or surpasses human doctors.

Another field profoundly affected by deep learning is natural language processing (NLP). Deep learning models, especially recurrent neural networks (RNNs) and more recently transformer architectures [15] (such as GPT), have transformed the way machines understand, generate, and translate human language. This has led to the development of advanced chatbots, real-time language translation tools, sentiment analysis, and text summarization systems. In the realm of business, deep learning-powered NLP models are used for customer support, automating repetitive tasks like answering frequently asked questions or analyzing customer feedback for insights.

Yet, despite the enormous progress and the huge potential of deep learning, the mathematical understanding of it deep learning is lacking far behind. Why does it work so well? Why does it sometimes fail miserably? How can we guarantee the convergence of training algorithms? What is the relationship between model capacity (e.g., number of parameters) and generalization? How do transformer architectures work so well for NLP tasks? What do hidden layers represent mathematically? How expressive are neural networks? Can we mathematically formalize fairness in AI systems? How can deep learning models be scaled efficiently? How can we interpret and explain deep learning models? These are just some of the essential questions surround deep learning for which so far we have no rigorous answers.

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.

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? Intriguing answers to some of those questions were given in several of the workshop talks, see Section 3.

2.1.1 Deep learning, frames, and neural collapse

Neural collapse is a phenomenon that emerges at the terminal phase of the training of deep neural networks (DNNs) [9]. The features of the data in the same class collapse to their respective sample means and the sample means exhibit a simplex equiangular tight frame (ETF) [12]. In the past few years, there has been a surge of works that focus on explaining why the neural collapse occurs and how it affects generalization. Since the DNNs are notoriously difficult to analyze, most works mainly focus on the unconstrained feature model. This model explains the neural collapse to some extent, it fails to provide a complete picture of how the network architecture and the dataset affect neural collapse. Indeed, while significant progress has been made in understanding neural collapse, several open research problems remain: Generalization and Test Collapse: While neural collapse is well-observed on the training set, its precise connection to generalization performance and whether a similar collapse occurs on the test set is still an active area of debate. Some

research suggests that strong test collapse is infeasible in practice, and that train and test collapse can even be anti-correlated in certain scenarios. Understanding this relationship is crucial for leveraging neural collapse to improve generalization. Most theoretical analyses of neural collapse have focused on simplified models (e.g., unconstrained features model, linear networks) and specific loss functions (e.g., MSE). Extending these analyses to more complex loss functions like cross-entropy (which does not always admit closed-form solutions and often yields an ETF structure for the last layer features when combined with feature normalization or weight decay) and understanding the role of bias terms in the collapse phenomenon is important. The original theory of neural collapse primarily focused on balanced datasets. Understanding how NC manifests and impacts performance in the presence of class imbalance is a significant challenge. Phenomena like "minority collapse," where minority class features become indistinguishable, have been observed, and characterizing the conditions and thresholds for such collapse is an ongoing area of research

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.

The interplay between data analysis and high-dimensional probability is fundamental in modern statistics and machine learning, as datasets increasingly feature numerous variables [16]. High-dimensional probability provides the theoretical underpinnings for understanding phenomena that arise when the number of dimensions approaches or exceeds the number of observations. Concepts like the "curse of dimensionality," where data becomes sparse and distances between points behave counter-intuitively in high dimensions, directly impact the effectiveness of data analysis techniques such as clustering, classification, and regression. Exploring these topics in depth formed another pillar of this workshop, as deepening our understanding of high-dimensional probability allows data analysts to select appropriate tools, interpret results accurately, and design more efficient algorithms for extracting meaningful insights from complex, high-dimensional datasets.

2.2.1 Multimodal data

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.

2.2.2 Advanced data denoising

The development of advanced denoising algorithms is becoming increasingly critical due to the vast amount of data generated and the complexity of the models that process it. As machine learning models, particularly

deep learning algorithms, are being applied to diverse fields such as medical imaging, autonomous driving, and natural language processing, the quality of the input data directly impacts the performance and accuracy of the system. Noise in data—whether it's visual, auditory, or textual—can significantly degrade model performance, leading to errors, inefficiencies, or even unsafe decisions. Therefore, creating robust denoising algorithms is paramount to ensuring that AI systems can operate with high precision and reliability. Several workshop talks reported about exciting progress in this area, see next section.

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 the 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.

Transformers have revolutionized the field of Artificial Intelligence, particularly in Natural Language Processing (NLP) and increasingly in computer vision and other domains. Despite their groundbreaking success, transformers present several intriguing mathematical challenges. Several talks address the mathematical challenges behind transformers, see Section 3.

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 [11]. 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.

2.5 Data manifolds, optimal transport, and Wasserstein space

As briefly discussed before, manifolds are particularly important in high-dimensional data and non-Euclidean geometry [8]. Many machine learning problems, such as dimensionality reduction or data generation, require understanding the underlying geometry of data, especially when it lives on a non-linear manifold rather than in flat Euclidean space.

Optimal transport provides a powerful and flexible way to compare probability measures, of all shapes: absolutely continuous, degenerate, or discrete [10]. This includes of course point clouds, histograms of fea-

tures, 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.

The connection between optimal transport (OT), manifolds, and machine learning is rich and multidimensional. Optimal Transport can be used in machine learning tasks where the data lies on a manifold, and one wants to account for the intrinsic geometry of the manifold while comparing distributions For example, OT is often used in applications like aligning datasets from different domains or time series data, where the data may be viewed as lying on different manifolds. By using OT in such scenarios, one can find an optimal match between data points while accounting for their manifold structure. In deep learning, the representations learned by neural networks often lie on some complex manifold. OT can be used in the loss functions to align representations across different domains or tasks (e.g., domain-invariant features) while respecting their geometric structure. While OT and the associated Wasserstein distance are powerful tools in machine learning, they are rather expensive to compute. For instance, computing the Wasserstein barycenter is known to be NP-hard.

For this reason, OT has been somewhat limited in practical applications, particularly in settings that demand scalable and efficient algorithms for tasks such as classification, dimension reduction, and generation. To overcome these limitations, Linearized Optimal Transport provides an effective framework that embeds measure-valued data into a Hilbert space, bridging the gap between optimal transport theory and machine learning so that measure-valued data can be processed using standard, out-of-the-box machine learning algorithms. By "linearizing" the problem, Linearized Optimal Transport offers both computational tractability and the ability to leverage classic techniques in linear spaces, making it highly appealing for practitioners who need tools that are both practical and efficient. Alexander Cloninger, in his talk, *Linearized Optimal Transport: Theory and Applications* presented exciting progress in this field. What is the potential of OT in areas such as fair data representations, where one needs to optimize the tradeoff between utility and some definition of fairness? How can we efficiently denoise data that reside on some (unknown) manifold? Can we detect and describe the data manifold based on data samples? These are some of the critical questions that have been tackled in this workshop, see Section 3 for more details.

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 Rene Vidal on *Semantic Information and Matching Pursuit Algorithms for Explainable AI*. There is a significant interest in developing ML algorithms whose final predictions can be explained in terms understandable to a human. Providing such an "explanation" of the reasoning process in domain-specific terms can be crucial for the adoption of ML algorithms in risk-sensitive domains such as healthcare. This has motivated a number of approaches that seek to provide explanations for existing ML algorithms in a post-hoc manner. However, many of these approaches have been widely criticized for a variety of reasons and no clear methodology exists in the field for developing ML algorithms whose predictions are readily understandable by humans. To address this challenge, Prof. Vidal proposes a method for constructing high performance ML algorithms which are "explainable by design". Namely, his method makes its prediction by asking a sequence of domain- and task-specific yes/no queries about the data (akin to the game "20 questions"), each having a clear interpretation to the end-user. He then minimizes the expected number of queries needed for accurate prediction on any given input. This allows for human interpretable understanding of the prediction process by construction, as the questions which form the basis for the prediction are specified by the user as interpretable concepts about the data. Experiments on vision and NLP tasks demonstrated the efficacy of Vidal's approach and its superiority over post-hoc explanations. Vidal's presentation led to lively interactions and many interesting and challenging follow-up questions.

Gal Mishne presented her work on *Low Distortion Embedding with Bottom-up Manifold Learning*. With the ubiquity of high-dimensional datasets in various biological fields, identifying low-dimensional topological

manifolds within such datasets may reveal principles connecting latent variables to measurable instances in the world. The reliable discovery of such manifold structure in high-dimensional datasets can prove challenging, however, largely due to the introduction of distortion by leading manifold learning methods. The problem is further exacerbated by the lack of consensus on how to evaluate the quality of the recovered manifolds. Prof. Mishne presented a novel measure of distortion to evaluate low-dimensional representations obtained using different techniques. Mishne additionally developed a novel bottom-up manifold learning technique called Riemannian Alignment of Tangent Spaces (RATS) that aims to recover low-distortion embeddings of data, including the ability to embed closed manifolds into their intrinsic dimension using a unique tearing process. Compared to previous methods, she showed that RATS provides low-distortion embeddings that excel in the visualization and deciphering of latent variables across a range of idealized, biological, and surrogate datasets that mimic real-world data. Low-dimensional structures were also the topic of Sui Tang's presentation *Solving Estimation Problems of Dynamical Systems by Exploiting Low-Dimensional Structures*. This talk explores the flourishing intersection of machine learning and differential equations, aiming to transform data into equations that are both predictive and insightful regarding the underlying systems that generated them. Driven by the need to exploit dynamical data sets in complex physical systems, Tang seeks to achieve inference with provable performance and to construct generalizable, interpretable predictive models.

In his talk, Shuyang Ling presented some intriguing new insights into the aforementioned phenomenon of neural collapse. Focusing on shallow ReLU neural networks, Shuyang tried to understand how the width, depth, data dimension, and statistical property of the training dataset influence the neural collapse. He provided a complete characterization of when the neural collapse occurs for two or three-layer neural networks. For two-layer ReLU neural networks, a sufficient condition on when the global minimizer of the regularized empirical risk function exhibits the neural collapse configuration depends on the data dimension, sample size, and the signal-to-noise ratio in the data instead of the network width. For three-layer neural networks, Shuyang showed that the neural collapse occurs as long as the first layer is sufficiently wide. Regarding the connection between neural collapse and generalization, he showed the generalization heavily depends on the SNR (signal-to-noise ratio) in the data: even if the neural collapse occurs, the generalization can still be bad provided that the SNR in the data is too low. These results significantly extend the state-of-the-art theoretical analysis of the neural collapse under the UFM by characterizing the emergence of the neural collapse under shallow nonlinear networks and showing how it depends on data properties and network architecture

Modern machine learning methods are revolutionizing what we can do with data — from TikTok video recommendations to biomarker discovery in cancer research. Yet, the complexity of these deep models makes it harder to understand what functions these data-dependent models are computing, and which features they learn to be important for a given task. In his talk *What did my deep network learn?*, Jeremias Sulam reviewed two approaches for turning general deep learning models more interpretable. The first one studied an unsupervised setting in the context of imaging inverse problems and showed how to design and train networks that provide exact proximal operators that approximate that of the (log) prior distribution of the data. The second approach switched to supervised classification problems for computer vision, where Jeremias did re-think the use of Shapley coefficients for black-box model explanations.

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.

As machine learning powered decision-making becomes increasingly important in our daily lives, it is imperative to strive for fairness in the underlying data processing. Shizhou Xu, in his talk *Fair Data Repre-*

*sentation for Machine Learning at the Pareto Frontier*proposed a pre-processing algorithm for fair data representation via which supervised learning results in estimations of the Pareto frontier between prediction error and statistical disparity. He then applied the optimal affine transport to approach the post-processing Wasserstein barycenter characterization of the optimal fair L_2-objective supervised learning via a pre-processing data deformation. Furthermore, he showed that theWasserstein geodesics from the conditional (on sensitive information) distributions of the learning outcome to their barycenter characterize the Pareto frontier between L_2-loss and the average pairwise Wasserstein distance among sensitive groups on the learning outcome. Shizhou presented several intriguing numerical simulations to underscore the advantages of his approach: (1) the pre-processing step is compositive with arbitrary conditional expectation estimation supervised learning methods and unseen data; (2) the fair representation protects the sensitive information by limiting the inference capability of the remaining data with respect to the sensitive data; (3) the optimal affine maps are computationally efficient even for high-dimensional data. The impact and timeliness of this framework is further underscored by the fact that a Data Study Group at the Turing Institute highlighted Shizhou's work in one of their reports.

Alberto Bietti, Courtney Paquette, and George A Kevrekidis dedicated their presentations to getting mathematical insight into modern deep learning architectures. Large language models based on transformers have achieved great empirical successes. However, as they are deployed more widely, there is a growing need to better understand their internal mechanisms in order to make them more reliable. These models appear to store vast amounts of knowledge from their training data, and to adapt quickly to new information provided in their context or prompt. Bietti studied how transformers balance these two types of knowledge by considering a synthetic setup where tokens are generated from either global or context-specific bigram distributions. By a careful empirical analysis of the training process on a simplified two-layer transformer, he illustrated the fast learning of global bigrams and the slower development of an "induction head" mechanism for the in-context bigrams. He highlighted the role of weight matrices as associative memories, and provided a theoretical analysis on how gradients enable their learning during training. In particular, he study the role of data-distributional properties. Paquette considered the solvable neural scaling model with three parameters: data complexity, target complexity, and model-parameter-count. She used this neural scaling model to derive new predictions about the compute-limited, infinite-data scaling law regime. To train the neural scaling model, she ran one-pass stochastic gradient descent on a mean-squared loss. She then derived a representation of the loss curves which holds over all iteration counts and improves in accuracy as the model parameter count grows. Paquette then analyzed the compute-optimal model-parameter-count, and identified 4 phases (+3 subphases) in the data-complexity/target-complexity phase-plane. The phase boundaries are determined by the relative importance of model capacity, optimizer noise, and embedding of the features. She furthermore derived, with mathematical proof and extensive numerical evidence, the scaling-law exponents in all of these phases, in particular computing the optimal model-parameter-count as a function of floating point operation budget.

Kevrekidis, in his talk, *Conformal Disentanglement with Autoencoder Architectures* addressed the problem of perspective synthesis and perspective differentiation. He introduced a neural network autoencoder framework capable of two key task in this context: it is structured to identify 'common' variables, and, making use of orthogonality constraints to define geometric independence, to also identify disentangled 'uncommon' information originating from the heterogeneous sensors. He demonstrated applications in several computational examples.

Several other talks highlighted the importance of a thorough mathematical theory in our understanding of deep learning. The intriguing interplay between neural networks and quantization was at the center of the talk *Compressing neural networks for faster inference: sparsity, quantization, and low-rank approximation* by Rayan Saab. He discussed recent advances in the compression of pre-trained neural networks using both novel and existing computationally efficient algorithms. The approaches he considered leverage sparsity, low-rank approximations of weight matrices, and weight quantization to achieve significant reductions in model size, while maintaining performance. He provided rigorous theoretical error guarantees as well as numerical experiments, paving the way for exciting new advances in this area.

Rebecca Willet spoke about *Learning Low-rank Functions With Neural Networks*. Neural networks are increasingly prevalent and transformative across domains. Understanding how these networks operate in settings where mistakes can be costly (such as transportation, finance, healthcare, and law) is essential to uncovering potential failure modes. Many of these networks operate in the "overparameterized regime," in

which there are far more parameters than training samples, allowing the training data to be fit perfectly. What does this imply about the predictions the network will make on new samples? That is, if we train a neural network to interpolate training samples, what can we say about the interpolant, and how does this depend on the network architecture? In her talk Rebecca described insights into the role of network depth using the notion of representation costs – i.e., how much it "costs" for a neural network to represent various functions. Understanding representation costs helps reveal the role of network depth in machine learning and the types of functions learned, relating them to Barron and mixed variation function spaces, such as single- and multi-index models.

Several exciting applications were at the center of several outstanding presentations. Image alignment is a central task in computer vision with diverse applications across many fields. The goal of rigid image alignment is to find a linear transformation, typically a rotation and a translation, that minimizes the discrepancy of one image to another. In many applications, the images being compared may be diffeomorphic or drawn from different underlying distributions. These images are commonly encountered in areas such as medical imaging, cryogenic electron microscopy (cryo-EM), and astronomy, and we refer to these as heterogeneous images. Rigid alignment of heterogeneous images presents an additional challenge because standard correlation methods often fail to provide meaningful alignments as they are sensitive to deformations in the images. It has been shown that the landscape induced by the Euclidean distance and the standard inner product are highly irregular in the context of density and image alignment. Yunpeng Shi presented a fast algorithm for rigid alignment of heterogeneous images based on a minimization of the sliced 2-Wasserstein distance. He showed that his method s robust to heterogeneous images, and scalable to large numbers of images. He combined transport metrics with the speed of fast Fourier methods to develop an algorithm that aligns two $n \times n$ sized images in $\mathcal{O}n \log n)$ operations. He showed theoretically and experimentally that his method gives accurate rotational alignments while being robust to translations and deformations of the images.

Convex relaxations for physics simulations were at the core of Yuehaw Khoo's presentation. She explored adaptations of semidefinite programming relaxations for solving many-body physics problems. Her approach transformed a high-dimensional PDE problem into a convex optimization problem, setting it apart from traditional non-convex methods that rely on nonlinear re-parameterizations of the solution. In the context of statistical mechanics, she demonstrated how a mean-field type solution for an interacting particle Fokker-Planck equation can be provably recovered without resorting to non-convex optimization. For quantum mechanical systems, she presented a similar technique to obtain the ground state of a quantum system and introduce a near-linear time algorithm for solving the convex program using hierarchical matrices.

Sparsity played a key role throughout the workshop, since sparsity in data science and AI significantly boosts efficiency by reducing memory usage and speeding up computations, making models more scalable for large datasets. Sparsity also improves model performance and generalization by acting as a regularizer, which effectively performs feature selection and prevents overfitting. This leads to more interpretable models. Bernhard Bodmann in his talk *Sparse recovery for linear combinations of heat kernels on graphs* addressed the challenge of how to identify individual terms in a superposition of heat kernels on a graph. He established geometric conditions on the vertices at which these heat kernels are centered and find bounds on the time parameter governing the evolution under the heat semigroup that guarantee a successful recovery. This result can be viewed as a type of deconvolution on a graph.

Manifold learning algorithms aim to map high-dimensional data into lower dimensions while preserving local and global structure. In this talk, In her talk *Low distortion embeddings with bottom-up manifold learning*, Gal Mishne presented a manifold learning framework that constructs low-distortion local views of a dataset in lower dimensions and registers them to obtain a global embedding. Motivated by Jones, Maggioni, and Schul (2008), LDLE constructs local views by selecting subsets of the global eigenvectors of the graph Laplacian such that they are locally orthogonal. The global embedding is obtained by rigidly aligning these local views, which is solved iteratively. Mishne's global alignment formulation enables tearing manifolds so as to embed them into their intrinsic dimension, including manifolds without boundary and non-orientable manifolds. We define a strong and weak notion of global distortion to evaluate embeddings in low dimensions. She showed that Riemannian Gradient Descent (RGD) converges to an embedding with guaranteed low global distortion. Compared to competing manifold learning and data visualization approaches, she demonstrated that LDLE achieves lowest local and global distortion on real and synthetic datasets.

Advanced data denoising was the topic of talks by Ronen Talmon, William Leeb, Boris Landa, and Hau-Tieng Wu. Ronen considered the fundamental problem of detecting the number of complex exponentials

and estimating their parameters from a noisy signal using the Matrix Pencil (MP) method. He introduced the MP modes and presented their informative spectral structure. He then showed theoretically that these modes can be divided into signal and noise modes, where the signal modes exhibit a perturbed Vandermonde structure. Leveraging this structure, he proposed a new MP algorithm, termed the SAMP algorithm, which has two novel components: (i) a new and robust model order detection with theoretical guarantees; (ii) an efficient estimation of signal amplitudes. Ronen showed empirically that the SAMP algorithm significantly outperforms the standard MP method, particularly in challenging conditions with closely- spaced frequencies and low Signal-to-Noise Ratio (SNR) values, approaching the Cramer-Rao lower bound (CRB) for a broad SNR range. Additionally, compared with prevalent information- based criteria, he showed that SAMP is more computationally efficient and insensitive to noise distribution.

Hau-Tieng Wu, in his talk *Manifold denoising* presented an efficient manifold denoiser based on landmark diffusion and optimal shrinkage under the complicated high dimensional noise and compact manifold setup. It is flexible to handle several setups, including the high ambient space dimension with a manifold embedding that occupies a subspace of high or low dimensions, and the noise could be colored and dependent. A systematic comparison with other existing algorithms on both simulated and real datasets was provided.

William Leeb studied the properties of a family of distances between functions of a single variable. These distances are examples of integral probability metrics, and have been used previously for comparing probability measures on the line; special cases include the Earth Mover's Distance and the Kolmogorov Metric. He examined their properties for general signals, proving that they are robust to a broad class of deformations. We also establish corresponding robustness results for the induced sliced distances between multivariate functions. Finally, he established error bounds for approximating the univariate metrics from finite samples, and proved that these approximations are robust to additive Gaussian noise. The results were nicelyillustrated in numerical experiments, which include comparisons with Wasserstein distances.

Detecting and recovering a low-rank signal in a noisy data matrix is a fundamental task in data analysis. Typically, this task is addressed by inspecting and manipulating the spectrum of the observed data, e.g. thresholding the singular values of the data matrix at a certain critical level. This approach is well established in the case of homoskedastic noise, where the noise variance is identical across the entries. However, in numerous applications, the noise can be heteroskedastic, where the noise characteristics may vary considerably across the rows and columns of the data. In this scenario, the spectral behaviour of the noise can differ significantly from the homoskedastic case, posing various challenges for signal detection and recovery. To address these challenges, Boris Landa developed an adaptive normalization procedure that equalizes the average noise variance across the rows and columns of a given data matrix. His proposed procedure is data-driven and fully automatic, supporting a broad range of noise distributions, variance patterns and signal structures.

Randomized algorithms play a key role in data science and large-scale optimization [7, 13, 5]. Talks by Jason Altschuler, Nicholas Marshall, and Robert Webber reported on amazing new developments in this important area. Webber investigated two randomized preconditioning techniques for solving kernel ridge regression (KRR) problems with a medium to large number of data points, and it introduces two new methods with state-of-the-art performance. The first method, RPCholesky preconditioning, accurately solves the full-data KRR problem in $\mathcal{O}(N^2)$ arithmetic operations, assuming sufficiently rapid polynomial decay of the kernel matrix eigenvalues. The second method, KRILL preconditioning, offers an accurate solution to a restricted version of the KRR problem involving $k \ll N$ selected data centers at a cost of $\mathcal{O}((N+k2)k\log k)$ operations. The proposed methods solve a broad range of KRR problems, making them ideal for practical applications. Marshall studied the effect of adding geometrically smoothed momentum to the randomized Kaczmarz algorithm, which is an instance of stochastic gradient descent on a linear least squares loss function. He proved an intriguing result about the expected error in the direction of singular vectors of the matrix defining the least squares loss. He then presented several numerical examples illustrating the utility of our result and pose several questions.

Gradient descent is a key ingredient of many AI algorithms [14]. Can we accelerate the convergence of gradient descent without changing the algorithm – just by optimizing stepsizes? Surprisingly, the answer is yes, as Jason Altschuler explained in his very inspiring presentation. He showed that for separable convex optimization, random stepsizes fully accelerate Gradient Descent. Specifically, using inverse stepsizes i.i.d. from the Arcsine distribution improves the iteration complexity from $O(k)$ to $O(k^{1/2})$, where k is the condition number. No momentum or other algorithmic modifications are required. The starting point is a conceptual connection to potential theory: the variational characterization for the distribution of stepsizes with

fastest convergence rate mirrors the variational characterization for the distribution of charged particles with minimal logarithmic potential energy. The Arcsine distribution solves both variational characterizations due to a remarkable "equalization property" which in the physical context amounts to a constant potential over space, and in the optimization context amounts to an identical convergence rate over all quadratic functions. A key technical insight is that martingale arguments extend this phenomenon to all separable convex functions. One can interpret this equalization as an extreme form of hedging: by using this random distribution over stepsizes, Gradient Descent converges at exactly the same rate for all functions in the function class. Jason's deep insights pave the way for various exciting new research directions in optimization and AI.

4 Outcome of the meeting

The workshop was attended by an enthusiastic audience of 33 (15 of which funded by BIRS/CMO and 18 self-funded) participants. Many of the participants were in early stages of their career. The workshop has brought together world leading experts at the intersection of applied harmonic analysis, machine learning, optimization, and signal processing to present recent developments and to foster new interactions. The direct interaction of mathematicians, statisticians, engineers, and computer scientists, made possible by this workshop, has made for an efficient intellectual feedback loop, which is central to achieving the urgently needed breakthroughs in machine learning and data science.

The program followed a well-tested and successful workshop format from our two previous workshops at CMO. Each day will consist of a mixture of talks by junior and senior researchers. The program featured one 1-hour talk per day, and an average of seven 30-minute talks per day, separated by coffee breaks and a long lunch break. With this format all participants were able to present their research, while making sure not to overload the schedule and leave ample time for discussions and unstructured interactions. The workshop timeline had a loose form to leave enough time for people to meet one-on-one at the conference site, while the presentation schedule encourages and allows junior participants and participants from underrepresented groups to present their work. The hour-long talks, at the beginning of each day, were overview talks by well chosen speakers on cutting-edge topics in the field. We also asked these speakers to highlight key open problems in their specific field. This allowed the participants to quickly learn about recent developments and it gave junior researchers a great opportunity to familiarize themselves with exciting mathematical problems. Wednesday afternoon was reserved for an excursion, that found enthusiastic reception. A very valuable aspect for young researchers at a conference was to have direct access to the world-leading experts in their field as well as communication and interaction among each other. This is often impossible to realize at large conferences, where keynote speakers often attend only for their talk and then depart soon after. But a small and intimate setting like the proposed workshop at CMO is perfect for this.

The friendly and personal setting for breakfast, lunch, and dinner at the conference site provided an inviting and informal environment that makes it easy even for more introvert participants to engage in conversations and discussions with senior colleagues. Our past workshops have shown that one of the most essential features is to have enough unscheduled time during the workshop, so that young researchers have plenty of time to actually interact with senior and mid-career scientists in their field. These extensive interactions provided one-on-one mentoring in a natural setting, which is often most productive.

The workshop was a tremendous success and many participants called it (one of) the best scientific meetings in the last decade. It initiated new research directions and collaborations and at the same time inspired work on solving some difficult problems at the intersection of applied harmonic analysis, data science, and machine learning.

Indeed, 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. Besides making progress on the topics discussed in detail in the previous sections, the workshop made great progress in the following domains:

- *Fostering communications between different focus points of research.* One main goal of this workshop has been to gather the leading international experts in the areas of applied harmonic analysis and data processing, and to create a fertile environment for the discussion of theoretical, numerical and applied aspects of this thriving research area. The workshop was very successful in initiating such

fruitful discussions and in starting new scientific collaborations. The workshop also provided excellent opportunities for junior scientists to interact closely with world leaders in the field.

- *New ideas for fundamental questions.* Since applied harmonic analysis has recently started to significantly impact the research area of data processing, it is now the right time to discuss fundamental questions in this rapidly evolving research direction. Each branch of the itemized list has its own fundamental open questions with various sidelinks to the other branches. This meeting has provided a unique opportunity to the major experts in this area for proposing and discussing novel ideas and research problems and suggesting solutions to those questions.
- *Manifest the future direction of the field.* Bringing together the world experts working in the area of applied harmonic analysis and mathematical data processing has enabled a thorough discussion of the strategic directions of investigation and a formulation of the open questions in this area.

References

[1] 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.

[2] D. L. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

[3] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

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

[5] Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[6] S. Mallat. Group invariant scattering. *Comm. Pure Appl. Math.*, 65(10):1331–1398, 2012.

[7] Per-Gunnar Martinsson and Joel A Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

[8] Marina Meilă and Hanyu Zhang. Manifold learning: What, how, and why. *Annual Review of Statistics and Its Application*, 11(1):393–417, 2024.

[9] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[10] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[11] 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.

[12] T. Strohmer and R. Heath. Grassmannian frames with applications to coding and communications. *Applied and Computational Harmonic Analysis*, 14(3):257–275, 2003.

[13] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *J. Four. Anal. Appl.*, 15(4), 2009.

[14] Yingjie Tian, Yuqi Zhang, and Haibin Zhang. Recent advances in stochastic gradient descent in deep learning. *Mathematics*, 11(3):682, 2023.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[16] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[17] John Wright and Yi Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.