Mathematical Foundations of Data Science: Titles and Abstracts

Ewout van den Berg, IBM

The Ocean Tensor Package

Practical applications in data science often require the processing of vast amounts of data using compute-intensive algorithms. In recent years, the use of GPUs and other accelerator devices has become widespread in reducing computational time and even in enabling algorithms that would otherwise be prohibitively expensive. Software libraries are needed to take full advantage of these advanced compute devices, and in this talk we present the Ocean Tensor Package. The package provides a large number of basic operations for multi-dimensional arrays on CPU and GPU devices through a convenient interface. The package is designed to facilitate development of new modules, enable extensions to novel device types, as well as to support interoperability with other packages.

Simone Brugiapaglia, SFU

Less is more but structured is better: The benefits of structured sparsity in data science.

Sparsity is a fundamental principle that allows to describe complex objects by using a few degrees of freedom, leading to data compression and to reduced computational cost. In the last decades, this has led to important innovations in various fields of data science such as inverse problems in imaging, uncertainty quantification, statistical modeling, and machine learning.

The gain due to sparsity can be boosted when combined with additional a priori information regarding the structure of the object of interest. For example, sparsity patterns of wavelet coefficients of natural images are structured according to the wavelets subbands; moreover, smooth high-dimensional functions have sparse expansions with respect to orthogonal polynomials that are "downward closed" multi-index sets. In this talk, we will discuss the benefits of structured sparsity in different contexts of data science such as image processing, compressed sensing, high-dimensional function approximation, uncertainty quantification, and machine learning, by illustrating recent research results and presenting some related open problems.

Roger Donaldson, UBC

Finding Approximate Nearest Neighbours: A simple example of Locality Sensitive Hashing and it's application in a NoSQL database

Approximate Nearest Neighbours (ANN) methods attempt to find the element in a large corpus of vectors in R^d that is nearest to a query vector with high probability, as when d is large, an exhaustive search is slow. One such method for performing an ANN search, now relatively mature, is Locality Sensitive Hashing (LSH), whereupon distances between vectors are determined in a space nearly isometric to R^d, but where calculations are fast, convenient, or both. This talk describes a simple example of LSH. The mapped space in our simple example is interpretable, and nearness calculations can be performed on a NoSQL database in the same context in which text-based search is performed. We hope that this talk will be of particular appeal to researchers working with embeddings of complex objects (images, video, audio, text), where finding similar points in high-dimensional real spaces is of continued interest.

Navid Ghadermarzy, UBC

Near-optimal sample complexity for convex tensor completion

We study the problem of estimating a low-rank tensor when we have noisy observations of a subset of its entries. For a rank-r, order-d tensor in $R^{N\times N\times N\times \ldots \times N}$, the best sampling complexity achieved is $O\left(rN^{d/2}\right)$ which can be obtained by a tensor nuclear-norm minimization problem. This bound is significantly larger than $O(rdN)$, the number of free variables in a rank-r tensor. In order to close this gap, we introduce two norms: (i) the "M-norm", an atomic norm whose atoms are rank-1 sign tensors and (ii) "max-qnorm" which is a generalization of the matrix max-norm to tensors. We prove that when $r = O(1)$, we can achieve optimal sample complexity by constraining either one of two proxies for tensor rank, the convex M-norm or the non-convex max-qnorm. Furthermore, we show that these bounds are nearly minimax rate-optimal. This is joint work with Ozgur Yilmaz and Yaniv Plan.

Bamdad Hosseini, Caltech

Continuum limit of semi-supervised learning and spectral clustering on graphs

Graphical semi-supervised learning and clustering have attracted a lot of attention recently. Both problems are inherently geometrical and use very similar tools. Often, in both problems, a weighted graph is constructed that summarizes the similarities between pairs of points in the dataset. In this talk we use the graph Laplacian operator and study its spectral properties. In particular, we study the limit where the number of vertices (i.e., the size of the dataset) becomes large and the graph Laplacian converges to a differential operator that shares similar properties with the underlying discrete graph Laplacian. This suggests the definition of continuum limit analogues of semi-supervised learning and clustering problems that can be analyzed to gain insight into the discrete problems with large datasets.

Mark Iwen,MSU

Sublinear-Time Algorithms for Approximating Functions of Many Variables

The development of sublinear-time compressive sensing methods for signals which are sparse in Tensorized Bases of Bounded Orthonormal Functions (TBBOFs) will be discussed. These new methods are obtained from CoSaMP by replacing its usual support identification procedure with a new faster one inspired by fast Sparse Fourier Transform (SFT) techniques. The resulting sublinearized CoSaMP method allows for the rapid approximation of TBBOF-sparse functions of many variables which are too hideously high-dimensional to be learned by other means. Both numerics and theoretical recovery guarantees will be presented.
Coauthors: Bosu Choi (MSU), and Felix Krahmer (TUM).

Halyun Jeong, UBC

Linear convergence for variants of the randomized Kaczmarz algorithm

The classical randomized Kaczmarz algorithm is a popular method to solve overdetermined linear systems. It was shown to converge linearly in expectation under some mild condition. Due to the efficiency and simplicity of the algorithm, several variants of it have been proposed to accommodate linear inequality constraints, the magnitudes of linear measurements, solution sparsity, among others. Although these constraints are quite different in nature, the linear convergence for each variant has been either established or conjectured as well. In this talk, I will present these recent developments in the analysis of Kaczmarz algorithms.

Xiaowei Li, UBC

Concentration for Euclidean Norm of Random Vectors

We present a Bernstein's inequality for sum of mean-zero independent sub-exponential random variables with absolutely bounded first absolute moment. We use this to prove a tight concentration bound for the Euclidean norm of sub-gaussian random vectors and the concentration of sub-gaussian matrices on geometric sets. As an application, we discuss its implications for dimensionality reduction and Johnson-Lindenstrauss transforms.

Maxwell Libbrecht, SFU

Understanding the human genome through unsupervised machine learning

Despite having sequenced the human genome over fifteen years ago, much is still unknown about how it functions. With the advent of high-throughput genomics technologies, it is now possible to measure properties of the genome across the entire genome in a single experiment, such as measuring where a given protein binds to the DNA or what genes are expressed. However, the complexity and massive scale of these data sets--billions of base pairs with thousands of measurements each--pose challenges to their analysis. My research focuses on the development of new machine learning methods that address the challenges posed by genomics data sets.

I will focus on a method for combining probabilistic models with graph-based methods for semi-supervised learning. Graph-based based methods have been successful in solving many types of semi-supervised learning problems by optimizing a graph smoothness criterion. This criterion states that data instances nearby in a given graph are likely to have similar properties. A graph smoothness criterion cannot be directly incorporated into a generative unsupervised model because it is usually not clear what probabilistic process generated the data instances with respect to the graph, and incorporating the graph directly into a factorizable (i.e. time-series) model would break the model's factorizable structure, making exact inference methods like belief propagation intractable. This

method, called entropic graph-based posterior regularization (EGPR) provides a way to express a graph smoothness criterion in a probabilistic model by defining a regularization term on an auxiliary posterior distribution variable. We applied this approach to regulatory genomics data sets from the human genome, leading to the discovery of a new type of regulatory domain.

Hassan Mansour, Mitsubishi Electric Research Labs

Fused-Lasso Optimization and its Application to Radar Sensor Calibration

We derive an optimization algorithm for the fused-Lasso (L1 + TV) minimization problem. The fused-Lasso penalty was proposed as a generalization of the Lasso that is designed for learning classifiers of datasets that exhibit a natural ordering of their features. Our framework leverages tools that have been developed for non-smooth gauge minimization problems, and proposes efficient projectors onto the fused-Lasso penalty. We also present an application of fused-Lasso optimization in the context of blind calibration of sensors in distributed radar imaging.

Brian Macdonald, GreaterThanPlusMinus LLC

Optimization problems in professional sports

We give an overview of how data visualization, analysis, and optimization can be used within the National Hockey League and in the sports industry in general, in a variety of different contexts. We discuss how analytics can be used to assist a team's front office, coaching staff, and scouting department. We also discuss the kinds of data and optimization problems we encounter on the business side of the organization in departments like sales and marketing. Finally, we present an in-depth example of how data and optimization techniques can be used to help a league make decisions about realignment.

Abbas Mehrabian, McGill

Density estimation of mixtures of Gaussians and Ising models

Density estimation lies at the intersection of statistics, theoretical computer science, and machine learning. We review some old and new results on the sample complexities (also known as minimax convergence rates) of estimating densities of high-dimensional distributions, in particular mixtures of Gaussians and Ising models. Based on joint work with Hassan Ashtiani, Shai Ben-David, Luc Devroye, Nick Harvey, Christopher Liaw, Yaniv Plan, and Tommy Reddad.

Rayan Saab, UCSD

New and Improved Binary Embeddings of Data (and Quantization for Compressed Sensing with Structured Random Matrices)

We discuss two related problems that arise in the acquisition and processing of high-dimensional data. First, we consider distance-preserving fast binary embeddings. Here we propose fast methods to replace points from a subset of R^N with points in a lower-dimensional cube $\{\pm 1\}^m$, which we endow with an appropriate function to approximate Euclidean distances in the original space. Second, we consider a problem in the quantization (i.e., digitization) of compressed sensing measurements. Here, we deal with measurements arising from the so-called bounded orthonormal systems and partial circulant ensembles, which arise naturally in compressed sensing applications. In both these problems we show state-of-the art error bounds, and to our knowledge, some of our results are the first of their kind. This is joint work with Thang Huynh.

Mark Schmidt, UBC

Is Greedy Coordinate Descent a Terrible Algorithm?

There has been significant recent work on the theory and application of randomized coordinate descent algorithms, beginning with the work of Nesterov, who showed that a random-coordinate selection rule achieves the same convergence rate as the Gauss-Southwell selection rule. This result suggests that we should never use the Gauss-Southwell rule, as it is typically much more expensive than random selection. However, the empirical behaviours of these algorithms contradict this theoretical result: in applications where the computational costs of the selection rules are comparable, the Gauss-Southwell selection rule tends to perform substantially better than random coordinate selection. We give a simple analysis of the Gauss-Southwell rule showing that---except in extreme cases---it's convergence rate is faster than choosing random coordinates. Further, we (i) show that exact coordinate optimization improves the convergence rate for certain sparse problems, (ii) propose a Gauss-Southwell-Lipschitz rule that gives an even faster convergence rate given knowledge of the Lipschitz constants of the partial derivatives, and (iii) analyze proximal-gradient variants of the Gauss-Southwell rule.

Bruce Shepherd, UBC

Designing a network without knowing your traffic

Abstract TBC

Yifan Sun, UBC

Atomic pursuit: A gauge perspective

We investigate the problem of regularizing for atomic sparsity, popularized in compressed sensing and machine learning. It is well known that the conditional gradient method and Kelley's cutting plane method are equivalent under Lagrange duality, for this restricted class of problems. We extend these results to gauge duality, as well as illustrate close links with other greedy methods like Gauss-Southwell coordinate descent, and submodular optimization. Geometrically, these methods can be seen as producing a sequence of inscribing polytopes of the atomic norm ball. We show that these methods lead to efficient and scalable semidefinite optimization methods with low-rank solutions, producing a fast, scalable, SDP solver for phase retrieval and graph problems.