Banff International Research Station

for Mathematical Innovation and Discovery

Sparse Representations, Numerical Linear Algebra, and Optimization Workshop, October 5–10, 2014

MEALS

*Breakfast (Buffet): 7:00–9:30 am, Sally Borden Building, Monday–Friday
*Lunch (Buffet): 11:30–1:30 pm, Sally Borden Building, Monday–Friday
*Dinner (Buffet): 5:30–7:30 pm, Sally Borden Building, Sunday–Thursday
Coffee Breaks: In the foyer of the TransCanada Pipeline Pavilion (TCPL)
***Please scan your meal card at the host/hostess station in the dining room for each meal**

MEETING ROOMS

All lectures will be held in the lecture theater in the TransCanada Pipelines Pavilion (TCPL). An LCD projector, a laptop, a document camera, and blackboards are available for presentations.

SCHEDULE

Sunday

16:00	Checkin begins (Front Desk, Professional Development Centre, open 24 hours)
17:30–19:30	Buffet Dinner, Sally Borden Building
20:00	Informal gathering in 2nd floor lounge, Corbett Hall
	Beverages and a small assortment of snacks are available on a cash honor system

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
7:00–8:45	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast
8:45–9:00	Welcome	Breakfast	Breakfast	Breakfast	Breakfast
9:00–9:30	**Strohmer**	**Kressner**	**Balzano**	**Chadrasekaran**	**Hermann**
9:30–10:00	**survey CS**	**survey NLA**	**Stoll**	**Aravkin**	**van den Berg**
10:00–10:30	Coffee	Coffee	Coffee	Coffee	Coffee
10:30–11:00	**Tanner**	**Strakos/Liesen I**	**Discussion**	**Discussion**	**Vert**
11:00–11:30	**Temlyakov**	**Strakos/Liesen II**	**Wright/Yilmaz**	**Kutyniok/Saunders**	
11:30–11:40	Lunch	Group Photo	Lunch	Lunch	Lunch **Checkout**
11:40–1:00	Lunch	Lunch	Lunch	Lunch	Lunch **by noon**
1:00–1:30	Guided Tour	Lunch	Lunch	Lunch	Lunch
1:30–2:00	Guided Tour		Excursion		
2:00–2:30	**Willett**		Excursion		
2:30–3:00	**Ward**		Excursion		
3:00–3:30	Coffee	Coffee	Excursion	Coffee	
3:30–4:00	**Mixon**	**Friedlander**	Excursion	**Schmidt**	
4:00–4:30	**survey Frames**	**survey Opt**	Excursion	**Krahmer**	
4:30–5:00	**Casazza**	**Orban**	Excursion	**Roosta**	
5:00–5:30	**Saab**	**Cartis**	Excursion		
5:30–7:30	Dinner	Dinner	Dinner	Dinner	

Workshop participants are welcome to use BIRS facilities (BIRS Coffee Lounge, TCPL and Reading Room) until **3 pm Friday**, although participants must **checkout of the guest rooms by 12 noon**

Banff International Research Station

for Mathematical Innovation and Discovery

Sparse Representations, Numerical Linear Algebra, and Optimization Workshop, October 5–10, 2014

ABSTRACTS

Sasha Aravkin, IBM TJ Watson Research Center
Fast variational methods for matrix completion and robust PCA

We present a general variational framework that allows efficient algorithms for denoising problems (where a functional is minimized subject to a constraint on fitting the observed data). This framework is useful for both vector recovery (sparse optimization) and matrix recovery (matrix completion and robust principal component analysis). While our approach is quite general, its efficacy relies on efficient first-order solvers. For matrix completion and robust PCA, we incorporate new ideas in factorized and accelerated first-order methods into the variational framework to develop state of the art approaches for large-scale applications, and illustrate with both synthetic and real examples.

Laura Balzano, Dept of Electrical Engineering and Computer Science, University of Michigan
Local convergence of an algorithm for subspace identification with missing data

Low-dimensional linear subspace approximations to high-dimensional data find application in a great variety of applications where missing data are the norm, not only because of errors and failures in data collection, but because it may be impossible to collect and process all the desired measurements.

I will describe recent results on estimating subspace projections from incomplete data. I will discuss the convergence guarantees and performance of the algorithm GROUSE (Grassmannian Rank-One Update Subspace Estimation), a subspace-tracking algorithm that performs gradient descent on the Grassmannian. I will also discuss the relationship of GROUSE with an incremental SVD algorithm, and show results of GROUSE applied to problems in computer vision.

Coralia Cartis, Mathematical Institute, University of Oxford
Global efficiency of methods using probabilistic models for nonconvex optimization

Line-search methods are an important class of algorithms for unconstrained nonconvex optimisation that rely on approximately computing a local descent direction and computing a step along this direction in such a way that sufficient decrease is achieved. To ensure that sufficient decrease is possible, the direction has to satisfy certain requirements. Often in practical applications, particularly in large-scale ones, ensuring that the direction properties are always what is needed is prohibitively expensive or impossible. This may be because derivative information about the objective function is not available, or full gradient (and Hessian) evaluations are too expensive, or a model of the true objective function is too expensive to optimize accurately.

Here we present global convergence rates for a line-search method that is based on random models and directions whose quality is ensured only with certain probability. We show that as a function of the accuracy tolerance, the evaluation complexity of such a method is of the same order as if deterministic exact first-order models were used, and the probabilistic models only increase the complexity by a constant multiple. We particularise and improve these results in the convex and strongly convex case. We also analyse a probabilistic cubic regularisation variant that allows approximate probabilistic second-order models and show improved complexity bounds compared to probabilistic first-order methods.

(Joint work with Katya Scheinberg, Lehigh University.)

Pete Casazza, Dept of Mathematics, University of Missouri
Phase retrieval by projections

The problem of recovering a vector from the absolute values of its inner products against a family of measurement vectors has been well studied in mathematics and engineering. A generalization of this phase retrieval problem also exists in engineering: recovering a vector from measurements consisting of norms of its orthogonal projections onto a family of subspaces. There exist semidefinite programming algorithms to solve this problem, but much remains unknown for the more general case. Can families of subspaces for which such measurements are injective be completely classified? What is the minimal number of subspaces required to have injectivity? How closely does this problem compare to the usual phase retrieval problem with families of measurement vectors? We answer or make incremental steps toward these questions. We provide several characterizations of subspaces that yield injective measurements, and through a concrete construction we prove the surprising result that phase retrieval can be achieved with $2M-1$ projections of arbitrary rank.

Venkat Chandrasekaran, Computing and Mathematical Sciences & EE, Caltech
Relative entropy relaxations for signomial optimization

Signomial programs (SPs) are optimization problems consisting of an objective and constraints specified by signomials, which are sums of exponentials of linear functionals of a decision variable. SPs are non-convex optimization problems in general, and instances of NP-hard problems can be reduced to SPs. We describe a hierarchy of convex relaxations to obtain successively tighter lower bounds of the optimal value in SPs. This sequence of lower bounds is computed by solving increasingly larger-sized relative entropy optimization problems, which are convex programs specified in terms of linear and relative entropy functions.

Our approach relies crucially on the observation that the relative entropy function—by virtue of its joint convexity with respect to both arguments—provides a convex parametrization of certain sets of globally nonnegative signomials with efficiently computable nonnegativity certificates via the arithmetic-geometric-mean (AM/GM) inequality. By appealing to representation theorems from real algebraic geometry, we demonstrate that our sequence of lower bounds converges to the global optimum for broad classes of SPs. Finally, we discuss numerical experiments that demonstrate the effectiveness of our approach.

(Joint work with Parikshit Shah.)

Michael Friedlander, Dept of Mathematics, UC Davis
Survey talk *Algorithms for sparse optimization*

Many applications in signal processing and machine learning need to solve optimization problems whose solutions are in some sense sparse. The aim is often to find the "best" compact representation of some object. Convex optimization plays a key role. The race is on to develop fast algorithms for problems that are deceptively simple, yet challenge our very best approaches. I will describe the fundamental building blocks, and survey some of the main research threads in sparse optimization.

Felix Herrmann, Earth and Ocean Sciences, UBC
Randomized methods in exploration seismology

Exploration seismology is a field of extremely large data sets ($O(10^{15})$ measurements), large sets of unknowns ($O(10^9)$), and hyperbolic PDE's, each of which challenges solutions of this PDE-constrained inverse problem. I will present two competing approaches that were developed in my group to overcome some of these challenges, namely techniques inspired by stochastic optimization and by compressive sensing. Both approaches use randomized dimensionality reduction techniques to reduce the number of PDE solves, but differ in how they control the errors. I will also talk about potential benefits of adaptive sampling.

(Joint work with my group SLIM, Ozgur Yilmaz, and Tristan van Leeuwen, and Mark Schmidt.)

Felix Krahmer, Institute for Computational and Applied Mathematics, University of Göttingen
Random shortcuts for linear transforms

We present a randomized algorithm for approximating matrix-vector multiplication. At the core of our proof is the quest for certain partially derandomized constructions of Johnson–Lindenstrauss projections. Namely, we require that the rows of the corresponding matrix are sampled from a set of polynomial size. To find such projections, we make use of approximate spherical designs. The resulting procedure is particularly geared toward computing dictionary representations.

(Joint work with Dustin Mixon, Air Force Institute of Technology.)

Daniel Kressner, École Polytechnique Fédérale de Lausanne
Survey talk *Sparsity and low rank in numerical linear algebra*

One of the key drivers of developments in numerical linear algebra has traditionally been the need for solving large-scale linear systems and eigenvalue problems, as they arise from finite difference or finite element discretizations of partial difference equations. The sparsity of the involved matrices has sparked the development of some of the most ubiquitous tools in scientific computing: Krylov subspace methods and sparse direct solvers. Recently, other types of data-sparsity have gained importance. This includes, in particular, matrices with low-rank structures such as hierarchical matrices and hierarchically semi-separable matrices. Exploiting (approximate) data sparsity in the desired solution is another important direction, which often requires blending linear algebra with techniques from other disciplines, in particular from optimization. The purpose of this talk is to give a survey of these recent developments and try to convey some intuition on their scope of utility.

Dustin Mixon, Air Force Institute of Technology
Survey talk *Frame theory: Applications and open problems*

Frame theory is the study of overcomplete bases, and applications can be found in multiple settings, such as time-frequency analysis, compression, and machine learning. Each application brings different criteria for evaluating a frame, and much of frame theory is devoted to designing frames that meet such specifications. I will review several applications of frame theory to motivate different specifications of particular interest (namely, tightness, incoherence, fast transforms, and signal model representation), and then survey the state of the art in the corresponding design problems.

Dominique Orban, GERAD and École Polytechnique de Montréal
Linear algebra for matrix-free optimization

When formulated appropriately, the broad families of sequential quadratic programming, augmented Lagrangian and interior-point methods all require the solution of symmetric saddle-point linear systems. When regularization is employed, the systems become symmetric and quasi-definite. They are indefinite but their rich structure and strong relationships with definite systems enable specialized linear algebra and make them prime candidates for matrix-free implementations of optimization methods. We explore various formulations of the step equations in optimization and corresponding iterative methods that exploit their structure.

Fred Roosta-Khorasani, Dept of Computer Science, UBC
Implicit matrix trace estimators: Gaussian estimator, rank, stable-rank, and majorization order

We are concerned with Monte-Carlo methods for the estimation of the trace of an implicitly given matrix A, where the matrix information is only available through matrix-vector products. The need to estimate the trace of implicit matrices arises in many applications, where A is often symmetric positive semi-definite (SPSD). Thus, theoretical studies of accuracy and efficiency of these methods are very important.

The standard approach for estimating the trace of an implicit matrix involves averaging the quadratic forms of A with random vector realizations from a suitable probability distribution. Among all possible distributions, we concentrate on the Gaussian distribution and derive new and improved theoretical results bounding the number of matrix-vector products required in order to guarantee a probabilistic bound on the relative error of the trace estimation. These new bounds indicate direct and surprising connections between the performance of the Gaussian estimator, the rank of the matrix, and its stable-rank. The connections will be discussed and further established using the concept of majorization order among vectors of eigenvalues.

Time permitting, numerical examples will be presented to demonstrate the success of such estimators in reducing the computational complexity of large-scale nonlinear least squares problems.

Rayan Saab, Dept of Mathematics, UC San Diego
Random encoding of quantized frame coefficients and quantized compressed sensing measurements

Frames generalize the notion of bases and provide a useful tool for modeling the measurement (or sampling) process in several modern signal processing applications. In the digital era, such a measurement process is typically followed by quantization, or digitization.

In the case of Sigma-Delta quantization of frame coefficients, we show that a simple post-processing step consisting of a discrete random Johnson-Lindenstrauss embedding of the integrated bit-stream yields near-optimal approximation accuracy as a function of the number of bits used. The result holds with high probability on the draw of the embedding, allows efficient reconstruction, and holds for a wide class of frames including smooth frames and random frames. (Joint work with Mark Iwen.)

We also show that if the same encoding scheme is applied to quantized compressed sensing measurements (with a different reconstruction scheme implemented using convex optimization), it also yields near-optimal approximation accuracy as a function of the bit-rate. (Joint work with Rongrong Wang and Ozgur Yilmaz.)

Mark Schmidt, Dept of Computer Science, UBC
Minimizing finite sums with the stochastic average gradient

We propose the stochastic average gradient (SAG) method for optimizing the sum of a finite number of smooth convex functions. Like stochastic gradient (SG) methods, the SAG method's iteration cost is independent of the number of terms in the sum. However, by incorporating a memory of previous gradient values, SAG achieves a faster convergence rate than black-box SG methods. Specifically, under standard assumptions the convergence rate is improved from $O(1/k)$ to a linear convergence rate of the form $O(p^k)$ for some $p < 1$. Further, in many cases the convergence rate of the new method is also faster than black-box deterministic gradient methods, in terms of the number of gradient evaluations. Beyond these theoretical results, the algorithm also has a variety of appealing practical properties: it supports regularization and sparse datasets, it allows an adaptive step-size and has a termination criterion, it allows mini-batches, and its performance can be further improved by non-uniform sampling. Numerical experiments indicate that SAG often dramatically outperforms existing SG and deterministic gradient methods, and that the performance may be further improved through the use of non-uniform sampling strategies.

Martin Stoll, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg
Preconditioning for PDE-constrained optimization with sparse controls

We study problems from PDE-constrained optimization in which controls with a certain sparsity are preferred. In addition to the classical sparsity term we include a directional sparsity term introduced recently. Our main goal is to consider the solution of the discretized systems that are at the heart of the non-smooth Newton scheme we employ. We focus on the development of preconditioners for the saddle-point systems and their use within Krylov subspace solvers. As the systems are of saddle-point form we are typically concerned with finding robust approximations to the relevant Schur complement. We illustrate the efficiency of our approximation on several examples.

(Joint work with Roland Herzog and Gerd Wachsmuth from TU Chemnitz, Germany.)

Zdeněk Strakoš and Jörg Liesen, Charles University in Prague and TU Berlin

Sparsity, local and global information in numerical solution of PDEs

In the finite element method (FEM) for discretizing partial differential equations (PDEs), the finite-dimensional piecewise polynomial approximation subspaces are generated using locally supported basis functions that typically vanish on all but a small number of elements determining the decomposition of the domain. The discretized operator is then represented by a possibly very large sparse matrix, and the sparsity is considered among the main advantages of the FEM discretizations. On the other hand, when iterative methods (such as Krylov subspace methods) are applied for solving the resulting algebraic problem, a transformation of the discretized problem historically called preconditioning is needed in order to assure fast convergence. Preconditioners incorporating coarse space information (such as multilevel preconditioners or domain decomposition techniques with coarse space components) often prove particularly efficient. This can be (at least partially) attributed to the fact that they handle in a natural way the global exchange of information between the parts of the domain.

With a little notation, the solution of the linear algebraic system $Ax = b$ can be symbolically written as $x = A^{-1}b$. The solution of the algebraic problem is given by the matrix-vector multiplication with the generally *dense* matrix A^{-1}, which conforms to possible global interactions between the parts of the domain described by the original mathematical model. If such interactions are strong, then it is very hard to approximate $x = A^{-1}b$ using the Krylov subspaces $\text{span}\{b, Ab, A^2b, \ldots, A^{k-1}b\}$, $k = 1, 2, \ldots$, where the individual generating vectors are formed by the repeated multiplication of the *sparse* matrix A. The sparsity of the discretized matrix A therefore does not automatically mean an advantage and the matter should be considered within the context of the original infinite-dimensional mathematical model.

This contribution concerns several questions about sparsity and the interplay between the local discretization and global algebraic computation. In particular, mostly using a restriction to second-order elliptic boundary value problems and the method of conjugate gradients, we interpret algebraic preconditioning as transformation of the discretization basis and simultaneous change of the inner product in the associated function space [1]. Moreover, we compare the distribution of the algebraic and discretization errors over the domain, and interpret the algebraic error as a possible change of the discretization basis [3], [2, Chap 5].

(Joint work with Josef Málek and Jan Papež from the Charles University in Prague and the Institute of Computer Science, Academy of Sciences of the Czech Republic.)

Acknowledgement: This work is being supported by the ERC-CZ project LL1202: *Implicitly constituted material models: from theory through model reduction to efficient numerical methods*, financed by the Ministry of Education, Youth and Sports of the Czech Republic.

[1] J. Málek and Z. Strakoš. *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs.* SIAM Spotlights Series, SIAM, Philadelphia, 2014 (in print).

[2] J. Liesen and Z. Strakoš. *Krylov Subspace Methods: Principles and Analysis.* Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013.

[3] J. Papež, J. Liesen, and Z. Strakoš. *Distribution of the discretization and algebraic error in numerical solution of partial differential equations.* Linear Algebra Appl., 449 (2014), 89–114.

Thomas Strohmer, Dept of Mathematics, UC Davis
Survey talk *Compressive sensing and beyond: challenges and opportunities*

At the mathematical heart of compressive sensing lies the discovery that it is possible to reconstruct a sparse signal exactly from an underdetermined linear system of equations and that this can be done in a computationally efficient manner via convex programming. I will briefly review the key ingredients of compressive sensing and its extensions, such as incoherence and randomness, various notions of sparsity, as well as associated basic convex relaxations. The main part of the talk is devoted to current challenges and possible opportunities of compressive sensing with an emphasis on the problems that are strongly related to the main theme of this workshop. These include the chasm between discrete and continuous models; self-calibrating compressive sensing and uncertainty mitigation; as well as fast and provable non-convex algorithms related to sparsity.

Jared Tanner, Mathematics Institute, University of Oxford
Conjugate gradient iterative hard thresholding for compressed sensing and matrix completion

Compressed sensing and matrix completion are techniques by which simplicity in data can be exploited for more efficient data acquisition. The design and analysis of computationally efficient algorithms for these problems has been extensively studied over the last 8 years. We present a new algorithm that balances low per iteration complexity with fast asymptotic convergence. This algorithm has been shown to have faster recovery time than any other known algorithm, both for small-scale problems and for massively parallel GPU implementations. The new algorithm adapts the classical nonlinear conjugate gradient algorithm and shows the efficacy of a linear algebra perspective to compressed sensing and matrix completion.

(Joint work with Jeffrey D. Blanchard (Grinnell College) and Ke Wei (University of Oxford).)

Vladimir Temlyakov, Steklov Institute of Mathematics, University of South Carolina
Greedy sparse approximation in Banach spaces

We study sparse representations and sparse approximations with respect to redundant dictionaries. We address the problem of designing and analyzing greedy methods of approximation. A key question is: How to measure efficiency of a specific algorithm? In answering this question we prove Lebesgue-type inequalities for the algorithms under consideration. A very important new ingredient of the talk is that we perform our analysis in a Banach space instead of a Hilbert space. It is known that in many numerical problems, users are satisfied with a Hilbert space setting and do not consider a more general setting in a Banach space. There are known arguments that justify interest in Banach spaces. Here we give one more argument in favor of consideration of greedy approximation in Banach spaces. We introduce a condition on a dictionary in a Banach space—a generalization of the RIP concept in a Hilbert space. We analyze the Weak Chebyshev Greedy Algorithm, which is a generalization of the Orthogonal Greedy Algorithm (Orthogonal Matching Pursuit) for Banach spaces.

Ewout van den Berg, IBM TJ Watson Research Center
A hybrid quasi-Newton projected-gradient method with application to Lasso and basis-pursuit denoise

I present a new algorithm for the optimization of convex functions over a polyhedral set. The algorithm is a hybrid of the spectral projected-gradient and quasi-Newton methods in which the type of step is determined at each iteration. A practical application of the framework is the Lasso problem, which also appears as a subproblem in the basis-pursuit denoise solver SPGL1. Other important applications that could benefit from the proposed algorithm include bound-constrained optimization and optimization over the simplex.

Jean-Philippe Vert, MINES ParisTech

New matrix norms for structured matrix estimation

Non-smooth convex penalties have gained popularity to estimate sparse models through convex optimization procedures, such as the lasso or the group lasso. I discuss two new matrix norms useful to estimate matrices with particular structures. First, I show that a norm proposed by Xiao, Zhou and Wu (2010) for orthogonal transfer is an atomic norm which in some sense is optimal for estimating matrices with orthogonal columns, and can be extended to estimate vectors with disjoint support. Second, I present the (k, q)-trace norm, a new convex penalty to estimate low-rank matrices with sparse factors, and show through an estimation of its statistical dimension and numerical experiments that it outperforms the L1 norm, the trace norms, and any of their combinations for sparse, low-rank matrix inference, in particular for sparse principal component analysis.

Rachel Ward, Dept of Mathematics, UT Austin

Completing any low-rank matrix, provably

Matrix completion results usually hinge on the assumption that the underlying matrix satisfies a restrictive structural constraint—known as incoherence—on its row and column spaces. In these cases, the subset of elements is sampled uniformly at random. We show that in fact any $n \times n$ matrix of rank r can be exactly recovered from as few as $O(nr \log^2 n)$ randomly chosen elements, provided this random choice is made according to a specific biased distribution based on leverage scores of the underlying matrix. Moreover, we show that this specific form of sampling is nearly necessary, in a natural precise sense; this implies that other perhaps more intuitive sampling schemes fail.

We further establish three ways to use the above result for the case when leverage scores are not known a priori: (a) a sampling strategy for the case when only one of the row or column spaces is incoherent, (b) a two-phase sampling procedure for general matrices that first samples to estimate leverage scores followed by sampling for exact recovery, and (c) an analysis showing the advantages of weighted nuclear/trace-norm minimization over the vanilla unweighted formulation for the case of non-uniform sampling.

(Joint work with Yudong Chen, Srinadh Bhojanapalli, and Sujay Sanghavi.)

Rebecca Willett, Electrical and Computer Engineering Dept, UW Madison

Tracking dynamic point processes on networks

Cascading chains of interactions are a salient feature of many real-world social, biological, and financial networks. In social networks, social reciprocity accounts for retaliations in gang interactions, proxy wars in nation-state conflicts, or Internet memes shared via social media. Neuron spikes stimulate or inhibit spike activity in other neurons. Stock market shocks can trigger a contagion of jumps throughout a financial network. In these and other examples, we only observe individual events associated with network nodes, usually without knowledge of the underlying dynamic network structure. Here we address the challenge of tracking how events within such networks stimulate or influence future events. We adopt an online learning framework well-suited to streaming data, using a multivariate Hawkes model to encapsulate autoregressive features of observed events within the social network. Recent work on online learning in dynamic environments is leveraged not only to exploit the dynamics within the underlying network, but also to track that network structure as it evolves. Regret bounds and experimental results demonstrate that the proposed method (with no prior knowledge of the network) performs nearly as well as would be possible with full knowledge of the network.

(Joint work with Eric Hall.)