OPTIMA 92

Mathematical Optimization Society Newsletter

Philippe L. Toint

MOS Chair's Column

September 15, 2013. Dear all, yes, time flies fast and this is my farewell column as Chair of the Mathematical Optimization Society since Bill Cook, whom I congratulate very sincerely, has become the new Chair on September 1st, 2013. Let me take it as an occasion to look back on three years in office, three years which have seen a thriving Society and many events in the community of researchers in optimization worldwide.

I think I concur with the general feeling in saying that the most obvious and satisfying of these events has been the ISMP 2012 Berlin. This symposium was at the same time a great meeting scientifically, a very satisfactory gathering of colleagues (and often friends) from all over the planet, and a superbly managed organization, whose talent even included summoning bright sunshine over Berlin for us all. The Paul Tseng memorial lectureship was also awarded during the symposium for the first time. Once more, many thanks to Martin and his fantastic team. But, of course, there were and are other successful optimization conferences supported by MOS such as the IPCOs and the ICCOPT in Lisbon. The community has indeed been far from lazy in setting up excellent scientific venues.

Another important event for the Society, which, to be honest, occurred just before I took office, was the change from Mathematical Programming Society (MPS) to Mathematical Optimization Society (MOS). Widely supported by the membership and by the Council, this change gave us better visibility and a clearer identification in the vast arena of research. My job in this, beyond supporting and preparing the idea from the start with Steve, has been to iron out the many little things that resulted from this, from legal to practical.

This naturally brings to mind the changes that occurred during my term in the not-so-visible administrative part of the Society. With the support of the Council, the administrative help contracted with SIAM was extended, for a better service to our members. It is a real pleasure to attest here to the professionalism of Arlette Liberatore who helped us for so many years, and more recently of Joanne Cassetti, who has replaced Arlette, now enjoying perpetual holidays … The service provided was always for the benefit of MOS and in a strict (but most friendly) independence from SIAM. This help was most precious (aside from prodding a too busy Chair for action when needed) in the internal management of the reform of membership dues, now allowing multi-year and life memberships, and also in the practical side of managing the Council elections.

In accordance with our status, a brand new MOS Council was duly elected and took office after the Berlin Meeting. It was a pleasure to work with the previous roster of officers, and this pleasure is again the rule in the open discussions with the new one. I also took to heart the re-establishment of the society's Publications Committee, whose input was crucial in the renewal of several editorial boards members and also, significantly, for discussing with Springer Verlag, our publisher, a better online access for the content of our journals.

All in all, this has been a relatively busy term, where the actions I could take for the service of the Society were made only possible by the help and support contributed by many of you: friends of the Executive Committee and of the Council(s), members of various committees, conference organizers, editors in chief of our journals and members of their editorial boards, webmaster, technical support, and also by the confidence of the society's members. I wish to express here my most sincere thanks to all. Thank you so much for helping me over these three years, and for allowing me, when passing the torch to Bill, to hand him over a healthy, lively, active and enthusiastic MOS.

Note from the Editors

Dear Optima readers, some of you may find this issue of Optima a bit unusual. Instead of a survey of some optimization topic, this time we have an article by Robert D. Nowak, Benjamin Recht, and Joel A. Tropp and a discussion column by Steve Wright all on the topics discussed at a recent workshop on Systems, Information, Learning, and Optimization, which was held at the University of Wisconsin in June 2013. This workshop's main focus was not optimization, however, as you can see from the articles, and hopefully agree, optimization plays a key role in many research areas discussed at this workshop. Moreover, we think that the optimization community will benefit from learning about some of those areas and issues. We hope you will enjoy the articles as well as some other supplementary materials and reports.

Sam Burer
Volker Kaibel
Katya Scheinberg
Optima editors

Contents of Issue 92 / September 2013

Robert D. Nowak, Benjamin Recht, and Joel A. Tropp

Report from the 2013 SILO Workshop

The 2013 SILO (Systems, Information, Learning, and Optimization) Workshop took place at the University of Wisconsin-Madison on June 17–19, 2013. The workshop was inspired by the weekly SILO seminar series held at the Wisconsin Institute for Discovery. SILO began as a way to build a community of mathematically minded researchers at the University of Wisconsin. Organizers Recht and Nowak recognized that there were mathematicians scattered across departments at UW who were all working on similar problems, even though these researchers communicated in very different languages. The goal of the SILO seminar is to break down the figurative silos raised by departmental organization and to connect researchers across campus to study foundations of the mathematics of information. SILO also has a third connotation as homage to the literal silos established on the agriculture campus at UW.

The study of a new, cross-disciplinary mathematics of information has been coalescing outside of UW as well, and the 2013 SILO workshop aimed to bring members of that community together to look forward to the new problems and challenges in this rapidly growing area and to identify threads of common inquiry. Bringing together a diverse group of researchers from computer science, engineering, statistics and mathematics, the aim of the 2013 SILO workshop was to ask, "What's next?". More precisely, what are the foundational research challenges that we must address in the mathematics of information?

After consulting with the workshop participants, five topical themes were selected to focus the talks and discussion:

1. Data (re)presentations
2. Foundations of Man-Machine Co-Processing Systems
3. Mathematics of Contemporary Computing Substrates
4. Dynamical Data Analysis
5. Inferential Complexity

A half-day session was dedicated to each topic. Each session consisted of one or two introductory talks followed by a panel discussion and group discussion. This article summarizes some of our findings and conclusions.

The workshop was organized by Robert Nowak (Engineering, UW-Madison), Ben Recht (Computer Science, UW-Madison) and Joel Tropp (Computing & Mathematical Sciences, Caltech). The workshop participants were Laura Balzano (UM-Ann Arbor), Constantine Caramanis (UT-Austin), Venkat Chandrasekaran (Caltech), John Doyle (Caltech), John Duchi (UC-Berkeley), Maryam Fazel (Univ. Washington), Anna Gilbert (UM-Ann Arbor), Al Hero (UM-Ann Arbor), Ali Jadbabaie (Penn), John Lafferty (Univ. Chicago), Per-Gunnar Martinsson (UC-Boulder), Michael Mahoney (Stanford), Mauro Maggioni (Duke), Deanna Needell (Claremont-McKenna), Pablo Parrilo (MIT), Sasha Rakhlin (Penn), Philippe Rigollet (Princeton), Justin Romberg (Georgia Tech), Lorenzo Rozasco (IIT/MIT), Katya Scheinberg (Lehigh), Devavrat Shah (MIT), Aarti Singh (CMU), Rachel Ward (UT-Austin), and Rebecca Willett (Duke). Also participating were Nigel Boston, Stark Draper, Jordan Ellenberg, Christopher Re, Sebastien Roch, Karl Rohe, Grace Wahba, and Stephen Wright from UW-Madison. The SILO Workshop was partially supported with generous support from the NSF, AFOSR, and ONR.

1 Data (re)presentations

The first topic discussed was how to represent data optimally for large-scale analysis tasks. For many data analysis problems, the primary challenge is to identify the proper representation or *features* in the data. Once we have done so, simple algorithms such as the SVM, Lasso, or nearest neighbors can solve (apparently) difficult tasks. For example, while it may be difficult to recognize faces based on raw pixel values, preprocessing the images to find eyes, noses, and mouths can make the subject identification task simpler. This session surveyed the current state of the art for feature engineering, exploring generic methods to discover the best features for varied analytics tasks.

Most of the recent work in this area has focused on *dictionary learning*. The idea here is that good features consist of representations of data that provide good reconstructions of the data and that represent the data parsimoniously. Classic examples of such representations include frames and wavelets, but data-driven methods such as k-means, k-subspaces, nonnegative matrix factorization, and sparse coding all fall within this rubric. The dictionary learning problem is naturally cast as a non-convex optimization problem: we seek to minimize the reconstruction error subject to a fixed description length. So, for example, to find a sparsifying dictionary of a data set, we would look for a matrix such that each data element could be represented as a sparse combination of a few of its columns.

Except in rare cases, such as PCA, these non-convex dictionary learning problems do not have efficient solutions. Moreover, it is difficult to provide guarantees that these learned dictionaries will be able to represent unseen data. There is also almost no mathematical analysis of the "learnability" of good representations (with the notable exception of a paper at last year's Conference on Learning Theory [14]). There was a strong consensus at the workshop that we need to build a firm theoretical foundation for dictionary learning and data representation in general.

There were several other questions discussed with regards to when we can learn good representations. Is it possible to reliably detect when a sparse dictionary for a dataset exists or is this decision problem computationally intractable? How can we validate when features are "good" for analysis? In most cases, features are designed without consideration of a specific analysis task, and we expect the same features to be good for classification, clustering, tracking, or whatever task we might have in mind. To what extent can features be targeted for specific objectives?

Some of the most active and exciting research in dictionary learning is building hierarchical dictionaries for data. Such hierarchical algorithms are called"deep learning" in the machine learning community [11]. These deep learning problems consist of nested, multistage optimization problems, and heuristics are used to find approximate local minima to these problems. Although the practitioners in this area claim that they can learn features with no domain knowledge or modeling, closer inspection reveals that they incorporate a substantial amount of prior knowledge into their algorithms. For example, in image features, most deep learning algorithms start by building dictionaries of small image patches. At small enough of a scale, simple Gabor filters are sufficient to sparsely reconstruct image patches. Moreover, the hierarchical structure in most deep learning architectures is strongly tied to the two dimensional geometry of images: it would fail completely if the pixels were randomly permuted. A major question is to understand how much of the benefit of dictionary and deep learning could be engineered with ideas from approximation theory and wavelets. Preliminary work by Mallat suggests that carefully designed wavelets can already compete with learned dictionaries [3]. Understanding how engineered wavelets can be enhanced to compete with more sophisticated techniques from machine learning looks to be an exciting direction of future study. Moreover, understanding how to pose tractable optimization problems to search for deep representations of data remains an intriguing open problem.

2 Foundations of Man-Machine Co-Processing Systems

People and computers are coupled in an increasingly complex system. Mathematical frameworks for modeling, analyzing, and optimizing human-computer interaction remain in their infancy. So far, mathematical research has focused on specific problems arising in application domains such as social media analysis, recommendation systems, e-commerce, and cognitive science, rather than tackling human-computer interaction holistically. How should human responses and behaviors be modeled? How can we understand how to control and monitor decentralized decisions, spread or prevent contagion, or understand shocks and vulnerabilities in tightly interconnected networks of human interaction? What representations best capture human judgements and preferences? Humans are slow and expensive, machines are fast and cheap. How can we optimize the symbiosis of man and machine?

Human input is integral to many information processing systems. Human subjects can provide expert judgments, annotations, labelings, rankings, or ratings. These inputs are used for a wide range of optimization tasks to improve system performance. While computational capabilities continue to grow and computing becomes less costly, human capabilities are relatively static, and the cost of human input is increasing. It is apparent that humans may be the bottleneck. Optimizing the use of humans in information systems is a major challenge.

Several design challenges for collaborative human-machine systems were put forward in this session. Are there notions of load balancing and/or impedance matching that might guide the optimization process? Can we develop models and algorithms that account for human latency, delay, accuracy, and responsiveness? How can we perform testing and assess the reliability of human-machine systems? Human psychology can play a central role in the efficacy of human-machine systems. Given that humans have limited attention spans, suffer from fatigue and calibration issues, and are susceptible to priming and other biases, models for human behavior are important. Mathematical psychology, educational testing, and economics offer starting points for further research, but new models are needed in the Big Data era. How do aggregates of human subjects function? Can the reliability or accuracy of different human subjects be estimated and factored into system operations?

Adaptively and sequentially optimizing human resources is a key component of human-machine systems. Using all information and data currently available, the objective is to automatically select the most informative tasks for human assistance. This optimization problem has been considered in many fields, ranging from statistical decision theory and machine learning to psychology and economics. Mathematically, this sort of problem has been formulated in terms of multi-armed bandits [4], Markov decision processing, active learning [5, 9], and signal processing [10]. In certain cases, optimal policies are known, but usually these are under restrictive and sometimes unrealistic assumptions. Computationally feasible and provably effective methods for optimizing human resources in general settings remains a relatively open problem. Quantifying the gains associated with adaptive and sequential schemes is also an important direction for research, since such schemes may be more complex to implement and more sensitive to modeling assumptions than nonadaptive approaches.

There are many kinds of human feedback. Numerical evaluations, scores, and ratings are notoriously difficult to calibrate, and so comparative judgements (e.g., comparisons between decision options or alternative models) are often favored. Comparative judgements are known to be more reliable and reproducible, and the comparative judgements of multiple humans are more easily aggregated. New theory and methods for incorporating comparative judgements into large-scale systems via crowd-sourcing is a potentially fertile area of research.

3 Mathematical Models of Contemporary Computing Substrates

Distributed computing and networked algorithms are all the rage, but do the usual mathematical models and constraints match the realities of contemporary infrastructures? As computing architectures and network structures change, it is imperative that our models for analysis continue to keep pace.

In this session, we began with a survey of modern computer hardware and networks. We looked at the basic models of the modern workstation and how these machines are typically networked in data centers. One concern that arose is the fact that many algorithms for distributed computation assume interconnection schemes that do not reflect best practices from industry. However, with new methods of virtualization and software-defined networks, it may indeed be possible to create very complex interconnection and communication schemes that do not currently exist in hardware. We examined the fact that communication-efficient algorithms are also power-efficient because data movement is energy intensive [7].

A basic issue is that most theoretical work does not account for multiple levels of network speed. For example, even on a multicore workstation, the time to move data from registers to cache is tens of thousands of times faster than the time required to read that information off an idle disk. In this regard, understanding distributed algorithms with multiple link speeds seems like a clear problem to explore. New optimization analyses need to do careful bookkeeping of computation, space, and time [1].

Along these lines, signal processing, optimization, and numerical analysis researchers could benefit from incorporating precomputed libraries into their algorithms. By making an initial investment (say, a day or a month) to precompute commonly used primitives, it may be possible to design algorithms that are substantially faster than current methods. Such pre-computation optimization has been studied in detail in database research, and a fruitful project would be to adapt this literature to numerical computation.

We also discussed how data itself is organized and the challenges in modeling large-scale networks such as social graphs. Common random models do not accurately match the statistics of real-world networks, yet much of our understanding of large networks comes from analyzing these random models [12]. Graph algorithms should reflect the statistics of real data, not the synthetic ones. Indeed, performance on random graph models tends to say very little about performance in practice.

Finally, we spent a long time discussing how it is imperative for mathematical researchers to interface with practitioners in the rapidly changing field of large-scale computation. Theorists do not necessarily want to work on problems that will be mapped into products in the next quarterly cycle, but our community must adapt our models of computation to reflect modern trends in computing. This cannot happen without constant interaction with systems engineers and big-data consumers in the sciences.

4 Dynamical Data Analysis

Online decision making, forecasting and prediction, control, and adaptivity are all intimately related, but they are treated by nearly disjoint communities (signal processing, control, machine learning, statistics). This session explored the rich potential for joining theoretical tools from these disciplines. We asked if it is possible to use techniques from machine learning and signal processing to help

dynamical systems adapt to stochastic environments? Is it possible to use techniques from dynamical systems to allow data processing systems to adapt to time-varying data? Does the control-theoretic perspective have anything to add to our understanding of streaming algorithms? Can we find a closed-loop theory that applies in information, statistics, and learning?

Recent research on online optimization has incorporated dynamical models to obtain enhanced versions of mirror descent [13]. How do these techniques compare with classic adaptive filters? Can these new methods give better theoretical guarantees for classic adaptive filtering methods? How do new techniques in sequential decision making and optimization relate to classic methods in optimal control theory? There is a disconnect here because control theorists have historically been interested in asymptotic rates of convergence, rather than those on a finite time horizon. Moreover, statistics has not always played a first-class role in optimal control theory. What can these two areas learn from each other?

Another way that data analysis can help dynamical models is by using machine learning to locally enhance control decisions [8]. Can we design predictors that use a modest number of measurements to steer a system toward an optimal configuration? How do modeling errors propagate? Tools from derivative-free optimization could play a role in establishing rates of convergence for this type of estimation scheme.

Diffusion processes on networks can be used to design distributed optimization, control, and decision making algorithms. There was some discussion about the concept of diffusion on higher-order complexes (as opposed to the edges of a network). This type of process might be able to tell us more about structures that exist (or do not exist) in a graph, and it may also have favorable convergence properties.

Complex networked systems, including communication, social, biological, and brain networks, are not stationary. Modeling the dynamics of networked systems and fitting models to observational data is a challenging problem. In some applications, the dynamics can be transient, with network topology and behavior changing abruptly.

Nonparametric estimation and detection methods are capable of automatically adapting to unknown spatial smoothness of signals. For example, wavelet-based methods have had tremendous success in signal and image processing. In comparison, theory and methods for adapting to unknown temporal dynamics are lacking. Can we develop signal processing and statistical inference methods that exploit hidden regularities in the dynamics of signals and systems?

5 Inferential Complexity

More data yields better inferences, but such quality always comes at a computational cost. Given a computational resource budget, is more data always helpful? Understanding the tradeoff between computational complexity and statistical accuracy (hence, inferential complexity) is a challenge of fundamental importance in the Big Data era, but little is currently known.

There have been recent attempts to answer this question using tools from theoretical computer science and statistical learning theory, but a general framework has yet to emerge. What are the sorts of mathematical tools that will allow us to build the appropriate bridges between computational complexity, mathematical statistics, and numerical linear algebra to understand the fundamental tradeoffs and hard limits in data analysis?

This session focused on trying to merge notions from mathematical statistics and theoretical computer science. mathematical statistics has well developed theory of tradeoffs between statistical accuracy and sample size, while theoretical computer science has well developed theory of tradeoffs between solution accuracy and computational complexity. In principle, these tradeoffs should be unified, and we sought to understand the cases where we could compute a Pareto curve mapping the tradeoff between computation and statistical confidence.

In this discussion, we encountered an interesting multidisciplinary question about the interaction between adaptivity and instance optimality. As is often the case in optimization research, worst-case analysis tends to lead to very pessimistic bounds. At the same time, analyses and algorithms that consider properties of the particular instance often yield faster algorithms in practice. Tying in detailed properties of individual instances could help to refine trade-offs between computation and statistical accuracy.

A major focus of this session was the sparse PCA problem, which requests sparse principal components of a data matrix [2]. This problem is intimately related to the maximum clique problem in graph theory which is notoriously hard to solve in practice. Results about maximum clique demonstrate that sparse PCA must be difficult in interesting parameter regimes unless P is equal to NP. Sparse PCA was the only example we could devise where simple optimization heuristics are not competitive with combinatorial search. As a sharp contrast, the statistical bounds derived for standard sparse optimization via the lasso are within a constant factor of those derived for exhaustive combinatorial enumeration (see [6] for an extensive list of examples).

This session featured a somewhat heated discussion about the fragility of lower bounds. Although we commonly believe that we have a very robust notion of minimax lower bounds in statistics, we still make very restrictive modeling assumptions to achieve these bounds. In computational complexity, on the other hand, lower bounds are very rare, but that may be because we are trying to prove lower bounds for a very rich and expressive model of computation (i.e., the Turing machine). There was some skepticism as to whether we could ever get "true" lower bounds. Is there a reasonable model of computation for statistical estimation that will make it easier to obtain lower bounds? Would simply restricting the model of computation to common algorithmic tasks like linear algebra or convex optimization allow us to develop a better understanding of trade-offs? Our discussion concluded that the best approach may be to trace out Pareto frontiers on an instance-by-instance basis.

One point of agreement is that constants matter in this area, an issue that also arose in the session on computing substrates. In many cases, statisticians are only interested in the scaling behavior of the error rate as a function of the number of samples. But the other "constants" are incredibly important. There are many algorithms that perform very poorly in practice but have the correct optimal scaling with respect to the number of samples. A major challenge is to calculate precise constants.

6 Future Workshops

We believe the workshop was a success and hope to hold similar meetings in the future. Let us record a few of the lessons we learned in the planning and execution of this workshop. We intend to incorporate these observations at future events.

We believe that it is essential (and refreshing) to have few talks and many breaks. Active moderation of the panels is critical to getting the most interaction from the group, and it took us a few sessions before we became comfortable with this style of interaction. We felt as if the meeting improved as it progressed because we all figured out how to work with the modular format. Let us elaborate on these points.

The unstructured time at the meeting was very productive, and every break witnessed heated and highly technical conversations among the participants. The other feature of the meeting that seemed to be effective was that all five sessions had the same structure. We felt that, as a group, we became better at navigating the structure of the meeting with each subsequent session.

The introductory speakers had the daunting task of preparing lectures for our unusual workshop format. The speakers truly rose to the occasion. Our only misgiving, if anything, is that we still may have allotted too much time for introductory talks! It is challenging to give a lengthy presentation that does not focus on research that you have been actively engaged in. We believe that shorter introductory talks would help make sure that the overview is truly general.

At most meetings, the "moderator" is simply the person holding the "5 minutes" sign when a speaker talks for fifteen minutes longer than their allotment. In the SILO workshop, the moderators played an active role to guarantee that all participants had adequate time to contribute. They also helped to keep the conversation focused. We believe that the moderators could be even more proactive in the future.

We very much enjoyed the meeting, and are already looking forward to SILO 2!

Robert Nowak, Department of Electrical Engineering, University of Wisconsin, Madison, WI, USA. nowak@ece.wisc.edu
Benjamin Recht, Department of Electrical Engineering and Computer Science and Department of Statistics, University of California, Berkeley, CA USA. brecht@berkeley.edu
Joel A. Tropp, Department of Applied and Computational Mathematics, California Institute of Technology, Pasadena, CA USA. jtropp@cms.caltech.edu

References

[1] Grey Ballard, James Demmel, and Andrew Gearhart. Communication bounds for heterogeneous architectures. In *23rd ACM Symposium on Parallelism in Algorithms and Architectures (SPAA 2011)*, 2011.

[2] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. *Journal of Machine Learning Research*, 30:1046–1066, 2013.

[3] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *arXiv preprint arXiv:1203.1513*, 2012.

[4] Sebastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Proceedings of the National Academy of Sciences*, 5(1):1–122, 2012.

[5] Rui Castro and Robert Nowak. Minimax bounds for active learning. *IEEE Info. Th.*, pages 2339–2353, 2008.

[6] Venkat Chandrasekaran and Michael I Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.

[7] Samuel H Fuller, Lynette I Millett, et al. *The Future of Computing Performance: Game Over or Next Level?* National Academies Press, 2011.

[8] Peng Guan, Maxim Raginsky, and Rebecca Willett. Online markov decision processes with kullback-leibler control cost. In *American Control Conference (ACC), 2012*, pages 1388–1393. IEEE, 2012.

[9] Steve Hanneke. Rates of convergence in active learning. *Annals of Statistics*, 39:333–361, 2011.

[10] J. Haupt, R.M. Castro, and R. Nowak. Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory*, 57(9):6222 –6235, 2011.

[11] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2146–2153. IEEE, 2009.

[12] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[13] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proceedings of the Conference on Learning Theory (COLT)*, 2013.

[14] Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

Stephen J. Wright

Remarks on Optimization in SILO

I was able to attend the SILO Workshop only by video hookup during the wee hours of the Australian morning. My biased sample of the live proceedings (and a later study of the slides for the introductory talks) confirms the organizers' opinion that the meeting was highly successful and that it highlighted some of the most exciting current research in data analysis and learning.

As an optimizer who has been marginally involved in these fields for some time, I was asked to make some remarks on SILO issues from the optimization perspective. I'll start with some background, then discuss the optimization issues that arise in data analysis and learning, and the ways in which optimization research (past and present) addresses these issues.

Data analysis can be defined broadly as the extraction of knowledge from data. Machine learning is similar in scope, but emphasizes the use of the knowledge to make predictions about other, similar data. These areas are highly interdisciplinary, drawing on statistics, information theory, signal processing, and computer science (artificial intelligence, databases, architecture, and systems). Optimization too is key. Not only is it embedded into many aspects of data analysis and learning (as discussed below), but it also plays a familiar role in turning the knowledge thus gained into good decisions.

Interest in data analysis and learning has grown because of the buzz surrounding "big data". A feature article in the New York Times Magazine (11 Feb 2012), quoted by Michael Mahoney in his SILO talk, opines that "(big data) opens the door to a new approach to understanding the world and making decisions". The scientific, social, and economic implications of big data will take years to fathom, and it may not live up to the hype, but the potential is clearly present for major impacts across many fields.

Important big data application problems are found in speech, language, and text processing (e.g., speech recognition, machine translation); image and video processing (e.g., denoising/deblurring and medical imaging); biology and bioinformatics (e.g., identifying genomic and evironmental risk factors for diseases); feature identification in geographical and astronomical images; and many other areas. As we discovered recently, U.S. government agencies have been busy solving big-data problems of their own, analyzing surveillance data from telephone and email communications.

The nature of the analysis differs across these applications, as does the use that is made of the extracted knowledge. Nevertheless, some powerful unifying themes can be identified. One theme is the prevalence of regression and classification problems. Given many items of data and an output or label associated with each item, can we learn a function that maps the data to its corresponding output? This function can then be applied to future, unknown items of data and used to predict the output. By parametrizing the function appropriately and applying statistical principles (for example, expressing the likelihood of the observations as a function of the parameters) such problems can be formulated as optimization problems. A process of this type leads to the familiar least-squares problem, and the only slightly less familiar robust regression, logistic regression, and support vector machine (SVM) formulations. (A common version of the latter is a structured convex quadratic program, to which many

optimization methods have been applied during the past 15 years.) Many formulations have partially separable objectives, a consequence of the fact that the data set has many items of the same structure to which the same transformations and measures are applied. Algorithms of stochastic and incremental gradient type have thus become extremely popular. Each iteration of these methods requires only a small, randomly selected subset of the data, using this sample to form an unbiased estimate the full objective gradient. These methods can be applied also to streaming data, provided we assume that the order of arrival of data items is random. Stochastic gradient methods date back to a 1951 paper of Robbins and Monro. They were studied independently by the machine learning and optimization communities for many years; forces have been joined in recent times. A particularly relevant property of stochastic methods is that they do not require evaluations of the objective, an operation that requires a complete sweep through the data set, and is therefore prohibitively expensive in some big data applications.

Another important theme is the identification of low-dimensional structure in high-dimensional data. Examples include finding a particular combination of base pairs in a genome (among an astronomical number of possible combinations) that indicate heightened risk of a disease, or finding a particular (possibly nonlinear) function of the pixel intensities in a picture of a digit, that makes it easy to identify the subject as being one of the digits 0 through 9. Two fundamental issues arise here. The first is one of *representation*, in which we seek ways to transform raw data into forms that facilitate more effective analysis. Deep learning – in which data is transformed by passing it though a layered neural network, resulting in output data that is easier to classify – is enjoying renewed popularity in speech and image processing. Optimization is used in the training of deep learning networks, in determining optimal values for the parameters that define the transformations at each layer of the network. Another way to address the representation issue is to choose a collection of basis elements (sometimes called "atoms") in high-dimensional space and define the low-dimensional structure in terms of a small subset of these elements. The basis can be predefined, or built up greedily or adaptively during the computation. Basis selection leads us to the second key issue: Formulation and solution of optimization problems that are tractable representations of the essentially intractable problem of low-dimensional structure identification. To explain: Consider the classical problem of finding the vector in R^n with $k \ll n$ nonzeros that minimizes a least-squares objective. A general algorithm would require investigation of all $\binom{n}{k}$ possible locations for the nonzeros, but compressed sensing shows us that when the least-squares objective has certain properties, a convex optimization formulation involving the ℓ_1 norm finds the solution. More generally, the challenge is to find regularization functions that can be included in the optimization formulation to induce the desired low-dimensional structure. The form of these functions depends, naturally, on the type of structure desired. As examples: The nuclear norm of a matrix tends to induce low rank in the solution of matrix optimization problems, and the use of the total-variation norm in image processing yields images with a natural quality – fields of constant color separated by sharp edges. Regularizations functions are often simple but nonsmooth. The study of formulation and solution of such problems is sometimes known as "sparse optimization."

Optimization formulations derived from Bayesian principles contain terms arising from prior assumptions about the knowledge hidden in the data. These terms often have similar forms to the regularization functions discussed above. Optimizers can leave the Bayesian vs. frequentist disputes to statisticians! Both approaches give rise to interesting optimization problems.

Partial separability and the widespread use of regularization are two typical characteristics of optimization problems in data analysis and learning. We mention several other ways in which these problems are unusual, by the standards of traditional optimization.

1. The objective functions often have a simple analytical form, making it easy to hand-calculate derivatives. (Indeed, it is argued that greater volumes of data make it possible to use less sophisticated models.)
2. Data scientists usually do not require a near-exact solution of the optimization problem, as the problem posed is often thought of as an empirical model (based on sampled data) of some underlying true objective. In fact, over-precise solution can lead to overfitting of the available data, at the expense of generalizability, that is, relevance of the solution to unseen data. In this sense, early termination of the optimization algorithm can be regarded as a form of regularization. The low-accuracy imperative is another reason for the success of stochastic gradient and first-order methods, which can sometimes find crude solutions rapidly.
3. Optimization formulations in these areas often contain simple scalar parameters, that trade off between different objectives, for example, between fitting the available data vs generalizability/regularization. The process of finding good values for these parameters is called "tuning." Often, the solution of the optimization model for a particular parameter is evaluated by some external criterion, such as its performance in predicting outputs for data items in a validation data set. The optimal parameter value is taken to be the one whose solution performs best on this criterion. Consequently, we need to solve not just one isolated problem, but rather a sequence of closely related problems, differing only in the choice of tuning parameters. Warm starting – using the solution for one value of tuning parameter as the starting point for a nearby value – has been applied with success. Moreover, techniques from derivative-free optimization can be used to traverse the space of tuning parameters, when the dimension is greater than one.
4. Data scientists are strongly interested in the theoretical complexity of optimization algorithms, such as different sublinear convergence rates (for example $1/\sqrt{k}$ vs $1/k$ vs $1/k^2$ in iteration number k) and dependence of complexity on the dimension of the data space. The level of interest would seem unusual to many optimizers, who are used to seeing only weak relationships between theoretical complexity and practical performance. Optimization complexity plays into the field of inferential complexity, which explores the tradeoffs between the statistical quality of a solution and the complexity of attaining it.

Many established optimization techniques, including some regarded as old-fashioned, have proved to be extremely useful in tackling data analysis and learning problems. Augmented Lagrangian methods, in particular the alternating direction method of multipliers (ADMM), are important in regularized formulations and as a basis for parallel methods. Accelerated first-order methods are popular because they can be extended easily to regularized objectives and require little extra work or storage than steepest-descent approaches. These methods introduce "momentum" terms into search directions to improve convergence rates, and are cousins of such old approaches as conjugate-gradient and heavy-ball. The prox-linear framework has proved useful for regularized formulations; LBFGS and inexact Newton methods have been adapted with much success to learning applications; and even the conditional-gradient method (sometimes known as "Frank-Wolfe") is enjoying a revival, as a way to find compact representations greedily. Coordinate relaxation, not taken very seriously by optimizers for some years, has been used

with success in support vector machines since the 1990s, and is being applied in other areas too. Duality has also proved to be an important tool. Duals are sometimes easier to solve and may (as in support vector machines) lead to reformulations with more powerful statistical properties. Primal-dual algorithms are efficient for some applications.

Computational systems issues – database systems, computation and memory architectures, parallel computing – also play a central role in big data. The interaction of optimization algorithms with systems is opening up new opportunities for research, for example in fast parallel asynchronous variants of stochastic gradient and coordinate descent. The possibility of using GPUs has piqued the interest of several resarchers since about 2008. They remain difficult to exploit for several reasons (including ease-of-use and memory transfer rates) but the potential payoff in computational efficiency is large, so they may yet hold interest in some contexts.

What of the future? Although we do not know how research priorities in SILO will evolve, we can say with confidence that optimization will continue to play an important role. It has become deeply enmeshed in many aspects of SILO; interest in optimization is running high among data scientists. New optimization formulations will continue to proliferate, each bringing its own particular challenges. It is not hard to imagine that optimization solvers will provide important middleware for general purpose data-analysis toolboxes, or that optimization technology will form some of the glue in "human-in-the-loop" systems for data analysis. Finally, new and increasingly complex computing substrates are rewriting the rules of computational cost and parallel processing. Optimization algorithms will need to be rethought and reanalyzed to exploit these new realities.

I close with several references. The report [1] presents a perspective on big data from leaders of the data science community. The recent edited volume [2] collects papers from on optimization for machine learning, written by researchers in both fields, and at their interface. Finally, I recommend perusal of the slides from the SILO Workshop, which illustrate the impressive variety and depth of research at the intersection of systems, information, learning, and optimization.

Stephen J. Wright, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, USA. swright@cs.wisc.edu

References

[1] National Research Council. *Frontiers in Massive Data Analysis*. National Academies Press, Washington DC, 2013.

[2] S. Sra, S. Nowozin, and S. J. Wright, editors. *Optimization for Machine Learning*. NIPS Workshop Series. MIT Press, 2011.

EURO Mini-Conference on Optimization in the Natural Sciences

February 5-9, 2014, Aveiro, Portugal. The conference aims at discussing the latest research in optimization and its applications in the natural sciences. The main objective of the meeting is to encourage the exchange of knowledge and recent achievements, and to establish collaborations among researchers and practitioners working in the areas of: Computer Science/Optimization/Statistics, and Physics/Chemistry/Biology/Medical Sciences.

Confirmed Invited Speakers

Leonid Bunimovich, Georgia Institute of Technology, Atlanta, USA. *Giuseppe Buttazzo*, Department of Mathematics, University of Pisa, Italy. *Aleksander Dudin*, Belarusian State University. *Michael Greenacre*, Department of Economics and Business, University Pompeu Fabra, Barcelona, Spain. *Georgui Smirnov*, University of Minho, Portugal. *Sergei Tabachnikov*, Pennsylvania State University, Department of Mathematics, USA.

Important Dates

Deadline for abstract submission: November 1, 2013. Notification of acceptance: November 15, 2013. Deadline for registration fee payment for inclusion in the abstract book: January 15, 2014.

Special Issue

Papers related to the abstract presented at EURO mini 2014 can be considered for peer-reviewed publication in a special issue of Optimization. Any registered author can submit the paper after conference in date to be given. Please read more about the Call for Papers for the Optimization journal Special Issue.

Organizing Committee

Co-chairs: Alexander Plakhov, Tatiana Tchemisova, and Adelaide Freitas. *Members:* Vera Afreixo, Isabel Brás, Paula Carvalho, João Pedro Cruz, Jorge Sá Esteves, Pedro Macedo, António Pereira, Ricardo Pereira, Paula Rama. University of Aveiro, Portugal.

Further Information

http://minieuro2014.web.ua.pt

17th Conference on Integer Programming and Combinatorial Optimization (IPCO XVII): Call for Papers

June 23–25, 2014, Bonn, Germany. The IPCO conference is a forum for researchers and practitioners working on various aspects of integer programming and combinatorial optimization. The aim is to present recent developments in theory, computation, and applications. The scope of IPCO is viewed in a broad sense, to include algorithmic and structural results in integer programming and combinatorial optimization as well as revealing computational studies and novel applications of discrete optimization to practical problems.

Authors are invited to submit extended abstracts of their recent work by November 15, 2013; see the submission guidelines for more information. The Program Committee will select the papers to be presented on the basis of the submitted extended abstracts. Contributions are expected to be original, unpublished and not submitted to journals or conferences with proceedings before the notification date (January 31, 2014). Papers violating these requirements will not be considered by the Program Committee.

During the conference, approximately 33 papers will be presented in single-track sessions. Each lecture will be 30 minutes long. The proceedings will be published as Springer Lecture Notes in Computer Science. They will contain extended abstracts of all accepted submissions. It is expected that revised and extended versions will subsequently be submitted for publication in appropriate journals. Each participant will receive a copy of the proceedings at the conference.

Important Dates

Submission deadline: November 15, 2013 (23:59 CET).
Notification date: January 31, 2014.
Conference: June 23–25, 2014.

Program Committee

Flavia Bonomo (Universidad de Buenos Aires), Sam Burer (University of Iowa), Gérard Cornuéjols (Carnegie Mellon University), Satoru Fujishige (Kyoto University), Michael Jünger (Universität zu Köln), Matthias Köppe (University of California, Davis), Jon Lee, chair (University of Michigan), Jeff Linderoth (University of Wisconsin), Jean-Philippe Richard (University of Florida), András Sebő (CNRS, Laboratoire G-SCOP, Grenoble), Maxim Sviridenko (University of Warwick), Chaitanya Swamy (University of Waterloo), Jens Vygen (Universität Bonn), David P. Williamson (Cornell University), Laurence Wolsey (Université catholique de Louvain).

Further Information

http://www.or.uni-bonn.de/ipco/

Mathematical Programming, Series B Special Issue on *Integer Programming Under Uncertainty*: Call for Papers

The journal Mathematical Programming, Series B invites submissions of papers to a special issue on integer programming under uncertainty. Potential topics include, but are not restricted to, novel exact and approximation algorithms for stochastic and robust discrete optimization problems; sampling, scenario generation, uncertainty description for these models; cutting planes and extended formulations for uncertain integer programs; integer programming-based approaches for chance constrained models; and stochastic combinatorial optimization under uncertainty. Also of interest are mathematical programming innovations that allow the practical solution of applications of integer programming under uncertainty in areas such as logistics, energy systems, supply chain management, finance, medicine, and engineering design.

Authors should submit their papers via www.editorialmanager.com/mapr/ and select article type "S.I.: Int Prog Uncert – Series B" for consideration in this special issue.

Papers will be refereed according to the standard of Mathematical Programming, Series A. Due to limits in page volume, we are requesting that all papers be submitted using MP style files, and conform to a maximum of 25 pages. The necessary LaTeX files can be downloaded at www.eng.cam.ac.uk/~dr241/MPB/MPB-Springer-latex-style-files.zip.

Deadline for submission of full papers: January 31, 2014. We aim at completing a first review of all submissions by June 15, 2014.

Additional information about the special issue can be obtained from the guest editors:

- Shabbir Ahmed, Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology 765 Ferst Drive, Atlanta, GA 30332, sahmed@isye.gatech.edu
- Jeff Linderoth, Professor, School of Industrial and Systems Engineering University of Wisconsin-Madison 1513 University Ave., Madison, WI 53706, linderoth@wisc.edu

IMPRINT

Editor: Katya Scheinberg, Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA. katyas@lehigh.edu ■ **Co-Editors:** Samuel Burer, Department of Management Sciences, The University of Iowa, Iowa City, IA 52242-1994, USA. samuel-burer@uiowa.edu ■ Volker Kaibel, Institut für Mathematische Optimierung, Otto-von-Guericke Universität Magdeburg, Universitätsplatz 2, 39108 Magdeburg, Germany. kaibel@ovgu.de ■ **Founding Editor:** Donald W. Hearn ■ Published by the Mathematical Optimization Society ■ Design and typesetting by Christoph Eyrich, Berlin, Germany. optima@0x45.de ■ Printed by Oktoberdruck AG, Berlin, Germany.