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 Munro. 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 sys-

tems 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.