SIAG/OPT Views and News

A Forum for the SIAM Activity Group on Optimization

Volume 24 Number 1 October 2016

Contents

Article

Evolution of randomness in optimization methods for supervised machine learning

Katya Scheinberg
Industrial and Systems Engineering Department
Lehigh University
Bethlehem, PA
USA
katyas@lehigh.edu
http://coral.ise.lehigh.edu/katyas

1 Introduction

Machine learning (ML) as a field is relatively new. Its beginning is usually attributed to the seminal work of Vapnik [1] from 1995, but it has grown at exponential speed in the past 20 years. One can say that ML is a combination of learning theory, statistics, and optimization and as such has influenced all three fields tremendously. ML models, while virtually unknown in the optimization community in 1995, are one of the key topics of continuous optimization today.

This situation is ironic, in fact, because after Vapnik's work the core of ML models shifted to support vector machines (SVMs)–a well-defined and nice problem that is a convex QP with a simple constraint set. On the other hand, today, the optimization problem at the core of many ML models is no longer as clearly defined and is not usually as nice. The main focus of our article is the evolution of the central optimization problem in ML and how different fundamental assumptions result in different problem settings and algorithmic choices, from deterministic to stochastic. We will consider the early SVM setting, which lasted for about a decade, where the problem was considered to be deterministic. In the following decade the perspective on ML problems shifted to stochastic optimization. And, over the past couple of years, the perspective has begun changing once more: The problem is again considered to be deterministic, but stochastic (or perhaps we should say randomized) algorithms are often deemed as the methods of choice. We will outline this evolution and summarize some of the key algorithmic developments as well as outstanding questions.

This article is based on one optimizer's perspective of the development of the field in the past 15 years and is inevitably somewhat biased. Moreover, the field of optimization in machine learning is vast and quickly growing. Hence, this article is by no means comprehensive. Most references are omitted, results are oversimplified, and the technical content is meant to be simple and introductory. For a much more comprehensive and technically sound recent review, we refer readers to, for example, [2].

2 Optimization models in machine learning

Optimization problems arise in machine learning in the determination of *prediction* functions for some (unknown) set of data-label pairs $\{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X} \subset \mathbf{R}^d$ and $\mathcal{Y}$ may contain binary numbers, real numbers, or vectors. Specifically, given an input vector $x \in \mathbf{R}^d$, one aims to determine a function such that when the function is evaluated at x, one obtains the *best prediction* of its corresponding label y with $(x, y) \in \{\mathcal{X}, \mathcal{Y}\}$. In *supervised learning*, one determines such a function by using the information contained in a set $\{X, Y\}$ of n known and labeled samples, that is, pairs $(x_i, y_i) \in \{\mathcal{X}, \mathcal{Y}\}$ for $i \in \{1, \ldots, n\}$. Restricting attention to a family of functions $\{p(w, x) : w \in W \subseteq \mathbf{R}^p\}$ parameterized by the vector w, one aims to find w_* such that the value $p(w_*, x)$ that results from any given input $x \in \mathcal{X}$ is *best* at predicting the appropriate label y corresponding to x. For simplicity we will focus our attention on binary classification, where labels y can take values 1 and -1. Note that the predictors are often called *classifiers*.

Our aim is to define an optimization problem to find the *best* classifier/predictor; however, we first need to define the measure of quality by which classifiers are compared. The quality (or lack thereof) of a classifier is typically measured by its *loss* or *prediction error*. Given a classifier w, a function $p(w, x) \in \mathbf{R}$, and a particular data instance (x, y), the classical $0 - 1$ loss is defined as

$$\ell_{01}(p(w,x), y) = \begin{cases} 0 & \text{if} \quad yp(w,x) > 0 \\ 1 & \text{if} \quad yp(w,x) \le 0. \end{cases}$$

Thus, this loss outputs a 0 if the classifiers made a correct prediction or 1 if it made a mistake. Note that $\ell_{01}(p(w, x), y)$ is not a convex function of w regardless of the form of $p(w, x)$.

Since the data for which the prediction is needed is not known ahead of time, we assume that (x, y) is randomly selected from $\{\mathcal{X}, \mathcal{Y}\}$ according to some unknown distribution.[1] The quality of a classifier w can be measured as

$$\begin{aligned} f_{01}(w) &= \mathbb{E}_{(x,y)\sim\{\mathcal{X},\mathcal{Y}\}}[\ell_{01}(p(w,x),y)] \qquad (1) \\ &= P_{(x,y)\sim\{\mathcal{X},\mathcal{Y}\}}\{yp(w,x) > 0\}, \end{aligned}$$

which is the probability that $p(w, x)$ gives the correct prediction on data if the data is randomly selected from the distribution. This quantity is known as the *expected risk* and has a clear interpretation as a measure of the quality of a classifier. All that remains is to find a classifier that minimizes (1)!

Minimizing (1) is a stochastic optimization problem with unknown distribution. Such problems have been the focus of simulation optimization and sample average approximation techniques for a few decades [3–6]. Since the expectation in (1) cannot be computed, it is replaced by a sample average. Here, there is an important assumption to be made that will dictate which optimization approaches are viable: Can we sample from $\{\mathcal{X}, \mathcal{Y}\}$ indefinitely during the optimization process, or are we limited to a fixed set of sample data that is available a priori? The second option is a more practical assumption and indeed is usually the preferred choice in machine learning and dictates the difference between the optimization in ML and the traditional field of stochastic optimization. As we will see later, many optimization methods for ML are theoretically justified under the assumption of sampling indefinitely while being applied under the scenario of a fixed set of sample data.

Assume that we are given a data set $\{X, Y\}$: $X = \{x^1, x^2, \ldots, x^n\} \subset \mathbf{R}^d$ and $Y = \{y^1, y^2, \ldots, y^n\} \subset \{+1, -1\}$, with each data point (x^i, y^i) selected independently at random from $\{\mathcal{X}, \mathcal{Y}\}$. We consider a sample average approximation of (1):

$$\hat{f}_{01}(w) = \frac{1}{n}\textstyle\sum_{i=1}^{n}\ell_{01}(p(w,x_i),y_i). \qquad (2)$$

This is often referred to as the *empirical risk* of w.

We now have a pair of optimization problems: the *expected risk minimization* (the problem we actually *want* to solve),

$$w_* = \arg\min_{w\in W} f_{01}(w) = \mathbb{E}_{(x,y)\sim\{\mathcal{X},\mathcal{Y}\}}[\ell_{01}(p(w,x),y)], \qquad (3)$$

and the *empirical risk minimization* (the problem we *can* solve),

$$\hat{w} = \arg\min_{w\in W} \hat{f}_{01}(w) = \frac{1}{n}\textstyle\sum_{i=1}^{n}\ell_{01}(p(w,x_i),y_i), \qquad (4)$$

where $W \subset \mathbf{R}^p$. An important question of ML, considered by learning theory, is how the empirical risk minimizer $\hat{w}$ compares with the expected risk minimizer w_*. A vast amount of work has been done on bounding these quantities in learning theory; see, for example, [7–9]. Here we present simplified bounds, which help us convey the key dependencies. Given a sample set $\{X, Y\}$ drawn randomly and independently from the distribution $\{\mathcal{X}, \mathcal{Y}\}$, with probability at least $1 - \delta$, the following *generalization bound* holds for any w:

$$|\hat{f}_{01}(w) - f_{01}(w)| \le O\left(\sqrt{\frac{\mathcal{W} + \log(\frac{1}{\delta})}{n}}\right). \qquad (5)$$

Moreover, for $\hat{w}$ and w_*, the following *esimation bound* holds:

$$|f_{01}(\hat{w}) - f_{01}(w_*)| \le O\left(\sqrt{\frac{\mathcal{W} + \log(\frac{1}{\delta})}{n}}\right), \qquad (6)$$

[1] For notational simplicity, $\{\mathcal{X}, \mathcal{Y}\}$ will hence denote both the distribution of (x, y) and the set of all pairs (x, y), according to the context.

where $\mathcal{W} \in \Re$ measures the complexity of the set of all possible classifiers produced by $p(w, \cdot)$ for $w \in W$. The first bound (5) ensures that at the optimal solution $\hat{w}$, the expected risk represents the empirical risk whenever the right-hand side of (5) is small. The second bound (6) ensures that the quality of solution of (4) is almost as good as that of the best classifier from W. Thus, to learn "effectively," we need to balance the complexity $\mathcal{W}$ and the size of the data set to keep the right-hand sides of (5) and (6) small. For instance, in a "big data" setting, when n is very large, very complex models with large $\mathcal{W}$ and complex $p(w, \cdot)$, such as deep neural networks, can be effective. In small data cases, however, the complexity $\mathcal{W}$ and the model class may need to be tightly restricted. Using a larger, more complex, model class can reduce $\hat{f}_{01}(\hat{w})$ and $f_{01}(w_*)$, since it may expand the feasible sets of (4) and (3); however, $f_{01}(\hat{w})$ may be large, compared with $\hat{f}_{01}(\hat{w})$ and $f_{01}(w_*)$, if n is too small compared with $\mathcal{W}$. This situation is known as overfitting the data. Later we will discuss some measures of complexity and ways of controlling them.

We conclude from this discussion that if $\sqrt{(\mathcal{W} + \log(\frac{1}{\delta}))/n}$ is sufficiently small, then minimizing empirical risk by solving (4) results in a classifier that approximately minimizes the expected risk (3). However, several drawbacks to solving (4) arise. First, because of the nonconvexity and discontinuity of the $0-1$ loss ℓ_{01}, it is an NP-hard problem. Second, the loss function is insensitive to the amount of classification suffered by a classifier for a particular data point. Whether this is desirable often depends on the application, but it is believed that (Lipschitz) continuous loss functions may be more robust for learning. Hence, various convex loss functions are used in practice. For example, the "hinge" loss, used in SVMs, is a convex approximation of the $0-1$ loss:

$$\ell_h(p(w,x),y) = \begin{cases} 0 & \text{if } \; yp(w,x) \geq 1 \\ 1 - yp(w,x) & \text{if } \; yp(w,x) < 1. \end{cases}$$

This loss function is a nonsmooth, Lipschitz-continuous function and is guaranteed to be convex in w if $p(w,x)$ is linear in w. The other popular alternative is the "logistic loss," which is written as

$$\ell_g(p(w,x),y) = \log(1 + e^{-yp(w,x)}).$$

This is a smooth, Lipschitz-continuous function and is never zero. It grows nearly linearly when the error is large (i.e., $yp(w,x)$ is very negative), and it approaches zero when there is no error and $yp(w,x)$ is large and positive. Like the hinge loss, this loss is convex over w provided $p(w,x)$ is linear in w.

Hence, in general, given a convex function $\ell(p(w,x),y)$, measuring the *loss* incurred when the label is determined by $p(w,x)$ while the true label is y gives two new optimization problems:

$$w_* = \arg\min\nolimits_{w \in W} f(w) = \mathbb{E}_{(x,y)\sim\{\mathcal{X},\mathcal{Y}\}}[\ell(p(w,x),y)], \quad (7)$$

called the *expected loss minimization* (the problem we want to solve), and

$$\hat{w} = \arg\min\nolimits_{w \in W} \hat{f}(w) = \frac{1}{n}\sum\nolimits_{i=1}^{n} \ell(p(w,x_i),y_i), \quad (8)$$

called the *empirical loss minimization* (the problem that can be solved). Fortunately, one can establish bounds similar to (5) and (6) for f, w_*, and $\hat{w}$ defined above using convex loss functions. The difference between the case of the $0-1$ loss we considered previously and the Lipschitz-continuous loss is the complexity measure, which is used in place of $\mathcal{W}$. We now briefly discuss complexity measures to illustrate the connection with the use of regularizers in learning.

Measures of complexity of $p(w, \cdot)$ and W are a central topic in learning theory and include well-known measures such as the VC dimension [1] (for $0-1$ loss) and Rademacher complexity [10, 11]. These measures depend on the set W and on the function $p(w,x)$, which is applied to compute the classifier output. For example, the most commonly used function $p(w,x)$ is the linear function $w^T x$. If $W = \mathbf{R}^d$, its VC dimension is known to be d. On the other hand, if for all $w \in W$, $\|w\| \leq R_W$ and for all $x \sim \mathcal{X}$, $\|x\| \leq R_X$, for some R_W and R_X, then the Rademacher complexity can be bounded by $O(R_X^2 R_W^2)$; hence, with probability at least $1-\delta$,

$$|\hat{f}(\hat{w}) - f(w^*)| \leq O\left(\sqrt{\frac{R_X^2 R_W^2 + \log(\frac{1}{\delta})}{n}}\right). \quad (9)$$

Numerous other bounds and results for the complexity measure $\mathcal{W}$ exist. In particular, the dependence of the VC dimension on the dimensionality of the feature space, d, at least partially justifies the dimensionality reduction that is so often employed, when the number of original features is large compared with the available data set. In particular, in the case of a linear classification, imposing a constraint on w that limits the number of nonzeros (i.e., $\|w\|_0 \leq t$ for some integer t) ensures that the VC dimension is bounded by t. Such a constraint turns the convex empirical loss minimization problem (8) into an NP-hard problem again. The standard technique is then to constrain the ℓ_1 norm of w as a relaxation of $\|w\|_0$; moreover, instead of adding an explicit constraint on $\|w\|_1$, a regularization term $\lambda\|w\|_1$ is added to the objective, in hopes of promoting the sparsity of $\hat{w}$ and thus the effective dimension of the data. Another popular regularization term is $\lambda\|w\|_2^2$, which naturally ties with the bound (9) since it tries to control R_W directly.

In summary, the classical ML optimization problem is formulated as

$$\min_{w \in \mathbf{R}^d} \; f_r(w) = \frac{1}{n}\sum\nolimits_{i=1}^{n} \ell(p(w,x_i),y_i) + \lambda r(w). \quad \text{(P)}$$

Here, the function r is a convex (potentially nonsmooth) *regularization* function whose presence is designed to control $\mathcal{W}$ and avoid overfitting. As mentioned, common choices are the ℓ_1 or ℓ_2 norm of the vector w. The choice of λ is not well understood from a theoretical standpoint. However, it is clear that large values of λ tend to result in smaller $\mathcal{W}$

but larger empirical loss $\hat{f}(\hat{w})$. In principle, one can formulate an optimization problem over λ that aims at minimizing $\hat{f}(\hat{w})$ and $\mathcal{W}$ together. Such a problem is referred to as *structural risk minimization* [7]; however, it is not tractable by itself. As a heuristic approach that is most commonly used, problem (P) is solved for various values of the hyperparameter $\lambda > 0$; then the best λ and the corresponding solution $\hat{w}$ are chosen based on the smallest *testing loss* (also known as the *validation error*). Specifically, a testing set $(X^t, Y^t) = \{(x_1^t, y_1^t), \ldots, (x_{n_t}^t, y_{n_t}^t)\}$ is allocated a priori (randomly selected from the same distribution $\{\mathcal{X}, \mathcal{Y}\}$), and the testing loss is simply

$$\frac{1}{n_t}\sum_{j=1}^{n_t} \ell(p(\hat{w}, x_j), y_j).$$

This approach provides an empirical way of finding the balance between the optimal empirical loss $\hat{f}(\hat{w})$ and the right-hand side of (5) and (6) to approximately minimize the expected loss $f(\hat{w})$.

3 Optimization methods for convex ML

The tremendous synergy between optimization and machine learning began in the late 1990s with the introduction of support vector machines. SVMs are a variant of (P), where $\ell(p(w, x_i), y_i)$ is the hinge loss, $p(w, x) = w^T\phi(x)$ with $\phi(x)$ being a mapping, and $r(w) = \|w\|^2$. The resulting optimization problem can be reformulated as a convex quadratic optimization problem. The power of SVMs lies in the fact that the feature vector $\phi(x)$ can be generated from the original data x by mapping x from $\mathbf{R}^d$ to a much higher-dimensional space $\mathbf{R}^p$, while the QP problem scales only with the number of data points n.

The interest in the optimization community was weak at first because, as far as the optimizers were concerned, convex QP had been solved. After all, between the active set methods and interior-point methods (IPM), most of the known QP instances in existence at that time could be easily handled. However, these methods turned out not to be well suited for large-scale SVM problems. Interior-point methods had been dismissed by the ML community initially because of their perceived high per-iteration complexity. In particular, since the QPs arising from the SVM model scale with the number of data points and many of them have dense Hessians, the complexity per iteration was seen as $O(n^3)$, and the required storage was seen as $O(n^2)$. In fact, closer inspection established that the complexity is $O(np^2)$ and the required storage is $O(np)$. Hence, if the feature dimension p is small (for example, in the case of linear SVM, when no mapping is employed, i.e., $\phi(x) = x$), then the per-iteration complexity of an IPM scales linearly with the size of the data [12]. The situation is more complex in the case of Kernel SVMs, where $\phi(x)$ may be a mapping into an infinite-dimensional space, but some dimensionality reduction can still be applied to reduce complexity [12, 13].

Active sets methods were not initially adopted by the ML community possibly because their theory does not guarantee convergence in polynomial time and their implementation is time consuming; however, a few methods have been successful in this application [14, 15]. Interior-point methods and active set methods are still being developed for large-scale SVMs [16], but they are not the main focus of the optimization in the ML community these days. The key reason perhaps lies in the ubiquitous claim that ML optimization problems *should not be solved accurately*, which is what IPMs and active set methods do. To an optimizer in the early 2000s not familiar with learning theory, this claim was puzzling. Why set up an optimization problem that one *wants* to solve inaccurately? Perhaps this is the wrong problem to solve? Of course, now we know that the answer is yes: problem (P) indeed *is* the wrong problem. The right problem is (3) or (7), for which (P) is just its surrogate.

To see why (P) does not need to be solved accurately, assume that we have obtained an approximate optimal solution to (P), $\hat{w}_\epsilon$, such that

$$\hat{f}(\hat{w}_\epsilon) \le \hat{f}(\hat{w}) + \epsilon.$$

Applying this and the bound (5) to $\hat{w}_\epsilon$ and $\hat{w}$ and using (6), we can easily show that

$$f(\hat{w}_\epsilon) \le f(w_*) + O\left(\sqrt{\frac{\mathcal{W} + \log(\frac{1}{\delta})}{n}}\right) + \epsilon. \tag{10}$$

From (10) one can see little benefit in reducing the optimization error ϵ much below $O\left(\sqrt{\frac{\mathcal{W}+\log(\frac{1}{\delta})}{n}}\right)$. But how can small optimization error be harmful? Why did some of the ML experts claim that problem (P) *should not* be solved accurately? This question has not been properly answered today, but some interesting insights have been put forth. While largely unsubstantiated by theory but supported by practice, the intuitive explanation is that solving (P) inaccurately by any algorithm reduces complexity $\mathcal{W}$ (by restricting W) and, hence, provides regularization. In other words, $\mathcal{W}$ acts inversely proportional to ϵ. While for most algorithms this has not been shown, some results connect early stopping of gradient descent with good generalization properties [17, 18].

A much more powerful justification for solving (P) inexactly but quickly is given, for example, in [19]. The idea is simple and beautiful. Recall that we want to solve (7), and let us select some accuracy ε and assume for the moment that our data set (X, Y) is not fixed but can be sampled indefinitely from $\{\mathcal{X}, \mathcal{Y}\}$. Assume that we apply an algorithm to solve (P) whose complexity of achieving accuracy ϵ for a given n is $c(n, \epsilon)$, which is increasing in n and ϵ^{-1}. Our goal is to compute an ϵ-accurate solution, $\hat{w}_\epsilon$ such that

$$f(\hat{w}_\epsilon) \le f(w_*) + \varepsilon, \tag{11}$$

for some given ε. Then, using (11), we should aim to have

$$O\left(\sqrt{\frac{\mathcal{W} + \log(\frac{1}{\delta})}{n}}\right) \le \frac{\varepsilon}{2} \quad \text{and} \quad \epsilon \le \frac{\varepsilon}{2}.$$

For fixed $p(w,\cdot)$ and W, with some given complexity $\mathcal{W}$, this implies that $n \geq O(\varepsilon^2)$ (note that some log factors are hidden, depending on the exact generalization bound that is used). Now let us consider the optimization algorithm and its complexity $c(n,\epsilon)$. Any algorithm that computes $\hat{f}(w)$ or the gradient (and perhaps Hessian) of $\hat{f}$ at each iteration has $O(n) = O(\frac{1}{\varepsilon^2})$ complexity per iteration, regardless of how fast it converges to the optimal solution. On the other hand, the stochastic gradient (SG) method does not compute $\hat{f}(w)$ and $\nabla\hat{f}(w)$; instead, at each iteration k it computes an unbiased estimate of $\nabla\hat{f}(w)$ based on a random sample x_k (or a subset of samples S_k) from $\{X,Y\}$:

$$\nabla\hat{f}_k(w) = \frac{1}{|S_k|}\sum_{i\in S_k}\nabla\ell_g(p(w,x_i),y_i) + \lambda\nabla r(w). \quad (12)$$

The step is then taken as $w^{k+1} = w^k - \alpha_k\hat{\nabla}f_k(w^k)$ with step size α_k following some rules (see the survey on the topic of SG methods and reference therein [2]). If the size of S_k is bounded by a constant, independent of n or k, then the SG method has $O(1)$ per iteration complexity. Thus, the SG method dominates any full-gradient-based method as long as it can converge to an $\frac{\varepsilon}{2}$-optimal solution in fewer than $O(\frac{1}{\varepsilon^2})$ iterations (since one computation of a full gradient is $O(\frac{1}{\varepsilon^2})$). This is indeed the case when (P) is strongly convex (e.g., when $r(w) = \|w\|^2$); in this case, the complexity of the SG method is $O(\frac{1}{\varepsilon})$ [19]. When (P) is not strongly convex, the complexity of a full-gradient-based method, such as an accelerated method, is $O(\frac{n}{\sqrt{\epsilon}}) = O(\frac{1}{\varepsilon^{2.5}})$. Hence, the SG method whose complexity is $O(\frac{1}{\varepsilon^2})$ is still dominant. In summary, while no clear evidence exits that solving (P) inaccurately is better than solving it accurately, one has a good theoretical reason to use a method that is slower in convergence but also slower in scaling with n.

The SG method has become a workhorse in machine learning and the default industry standard. However, it has drawbacks that are well known: Its convergence rate is slow, not only in theory but also in practice; it does not parallelize naturally; and it is sensitive to the choice of the step-size sequence α_k, which usually needs to be tuned for each application. In particular, the SG method tends to make great progress in the beginning when the true gradient is large, and then it stalls when the variance in the gradient estimates becomes significant compared with the size of the true gradient.

One common remedy is variance reduction, which is obtained by applying sample average approximation. In our setting of problem (P) this essentially means that the S_k sample set is chosen to be large. The question is how large? In recent work [20], the authors analyze a generic method for solving stochastic problems, such as (7), where S_k is increased according to the progress of the underlying deterministic algorithm. Since in a strongly convex case the gradient descent method with sufficiently small step sizes converges linearly, this implies that the size of S_k should grow exponentially with k. It then follows that for the total number of samples computed up to iteration K, $\sum_{k=1}^{K}|S_k|$ is proportional to the number of samples in the last iteration $|S_K|$, for any $K > 0$. Similar techniques of adaptively selecting S_k have been proposed in [21] and [2] for the strongly convex case. Results for nonconvex optimization will be mentioned when we talk about nonconvex ML models.

SG and the variance reduction techniques just discussed all assume that one can sample indefinitely from $\{\mathcal{X},\mathcal{Y}\}$. Only then can we expect to obtain an ε-accurate solution to (7) for any chosen ε. However, recall that in ML applications the sample set (X,Y) is usually given and fixed. The bad news is that no matter how accurately we solve the optimization problem (P), the solution will still be at most $\sqrt{(\mathcal{W}+\log(\frac{1}{\delta}))/n}$-accurate with respect to solving (7). However, the good news is that the SG method can be improved by exploiting the structure of (P) as a finite sum of n functions (plus a simple term). With this structure, several successful extensions of SG have been recently proposed: SAG [22], SAGA [23], SDCA [24], and SVRG [25]. SAG and SAGA, for example, rely on averaging the past stochastic gradients in a particular manner and, hence, accumulating more accurate gradient estimates. As a result, they enjoy the same convergence rate as do full-gradient methods but with improved constants. However, these methods require the storage of n past gradients. SVRG, on the other hand, does not need to store the gradients; however, it requires computing the full gradient every n iterations. On the remaining iterations it performs "corrections" using stochastic gradients and a fixed step size. As a result, SVRG also has the same convergence rate as the full-gradient method, and the complexity gain over the full-gradient method lies also (as in SAG and SAGA) in smaller constants. While these methods are still referred to as stochastic gradient methods, perhaps one might more appropriately call them randomized methods, since they use random steps and rely on the deterministic nature of (P).

These new randomized methods offer some balance between the slowly convergent and unstable SG method, which is theoretically optimal for ML problems, and fast-converging and robust, but costly, first- and second-order methods. Which methods are truly the most efficient depends on the application, computational environment, and data structure. Adapting SG methods to distributed environments and huge sparse data sets is a subject of a vast amount of ongoing work.

4 Deep learning and nonconvex models

While the use of convex models has led to great successes over the years and is still widespread in the ML community, *nonconvex* ML models, such as deep neural networks (DNNs), have become extremely popular because of their impressive predictive power on perceptual tasks such as speech and image recognition [26, 27].

We now describe a typical DNN. Given an input vector x, a DNN produces an output value $p(w,x)$ through a series of successive transformations. These transformations are computed by *neurons* that are arranged in one or more *layers*. Let a_j denote the input vector for layer j (with $a_j = x$ for the first layer) which is passed to each neuron in the layer.

Then each neuron indexed by, say, j_k produces an output $p_j(w_{j_k}, a_j)$, where p_j involves a nonlinear function, such as a componentwise sigmoid or a hyperbolic tangent function, and w_{j_k} denotes the set of parameters of j_kth neuron. The resulting vector consisting of $p_j(w_{j_k}, a_j)$ for all k in the jth layer then defines the input for the next layer, a_{j+1}. Note that different layers can have different numbers of neurons. Overall, the entire collection of parameters w needs to be determined, which is typically done by solving an optimization problem of the form (P). However, one can easily see that (P) is highly nonlinear and nonconvex even when the loss ℓ and regularization r are convex. Note also that the problem dimension can grow extremely quickly with the number of layers and the dimension of the input at each layer.

Hence, the optimization problem (P) for DNNs is nonconvex and high dimensional. Moreover, computation of the gradient of the objective of (P) is performed by *backpropagation*[2] [28] applied to every data point in X, which makes this process even more expensive than it is for convex ML models with large data sets. Second-order information is usually prohibitive to compute and store in the DNN setting; however, several matrix-free approaches have been proposed [29–31].

As in the convex case, while solving (P) may result in optimizing the empirical loss $\hat{f}(w)$[3], we are ultimately interested in minimizing expected loss $f(w)$, which is not computable. On the one hand, a more complex class of predictors, such as DNNs, contains some optimal predictor w^*, whose expected loss $f(w^*)$ is smaller compared with that of the predictors from a simpler (say linear) class. But does this imply that we can find such a powerful predictor in this complex class even when it exists? In other words, we need to bound $|\hat{f}(\hat{w}_\epsilon) - f(w^*)|$, where $\hat{w}_\epsilon$ is some approximate minimizer of (P). Since (P) is nonconvex, we cannot guarantee $|\hat{f}(\hat{w}_\epsilon) - \hat{f}(\hat{w})| \leq \epsilon$. Typically, one can guarantee $\|\nabla \hat{f}(\hat{w}_\epsilon)\|^2 \leq \epsilon$, which in the nonconvex case does not imply any bound on $|f(\hat{w}_\epsilon) - f(\hat{w})| \leq \epsilon$. Generalization bounds using the VC dimension of a DNN or other complexity measures can, in principle, provide bounds on $|\hat{f}(\hat{w}_\epsilon) - f(\hat{w}_\epsilon)|$ and $|\hat{f}(\hat{w}) - f(w^*)|$. That is, if the VC dimension of a DNN is small compared with the number of data points in X, then the expected loss of any w is not too different from its empirical loss, and the expected loss of the empirical loss minimizer $\hat{w}$ is not too different from the loss of the best possible w^*. However, even these bounds do not appear to be useful, since the VC dimension of a DNN with millions of parameters is very large and the training sets are often not sufficiently large to guarantee good generalization. In practice, regularizers such as $\lambda\|w\|_2^2$ are often used, but generalization bounds involving $\|w\|_2^2$ are unknown for DNNs. In summary, why one would want to try solve (P) for DNNs is unclear, given that it is difficult and it does not guarantee good predictors. The answer is unsatisfactory but compelling: Because it often works! This observation has been stated repeatedly for many machine learning applications and has resulted in a shift of focus from efficient optimization of convex ML models to the efficient optimization of nonconvex ML models.

With the new focus on nonconvex models, many optimization questions arise relating to what the objective function of (P) looks like. Empirical and theoretical evidence clearly shows that this function is very multimodal; hence, it is unreasonable to expect standard optimization methods to obtain a global minimizer. On the other hand, researchers have argued that local minimizers may be sufficient in terms of obtaining solutions with small empirical loss [32]. The argument is based on the analysis of random Gaussian functions in [33], which shows that the local stationary points tend to have low objective value if they are local minima and that the only stationary points with high objective value are likely to be saddle points. While the objective function (P) is not Gaussian, some numerical evidence was shown to support the same observation about its saddle points. Hence, avoiding saddle points and being content with local minima is seen as crucial in optimizing DNNs. One way of doing so, which has been known for a long time in continuous optimization, is to exploit negative curvature. On the other hand, without a $\lambda\|w\|_2^2$ regularization term, saddle points of the objective function of (P) for DNN are often second-order stationary points; hence, there is no negative curvature in the second-order Taylor model. If the $\lambda\|w\|^2$ regularization term is present, then all such saddle points become local minima, and thus the claim in [33] does not apply. In short, so far there seem to be no good understanding of the structure of (P) and how it should be exploited.

Currently, most DNNs are trained by using the SG method. The reasons for this choice are similar to those for the convex case, for example, low per-iteration cost and the ability to provide inexact solutions that have good generalization properties. However, little theoretical support for these properties exists in the DNN case. Also, while convergence of SG method to a local stationary point can be established for nonconvex problems, convergence rates are essentially unknown. In [34] a randomized version of SG is proposed, with a random stopping criterion that results in a random iterate X^i, for which $\mathbb{E}(\|\nabla f(X^i)\|^2) \leq \epsilon$ is achieved if the algorithm runs for at least $O(1/\epsilon^2)$ iterations. The expectation is taken over the random gradients and the random stopping of the algorithm. Essentially one needs to run the algorithm multiple times to obtain $\|\nabla f(X^i)\|^2 \leq \epsilon$ with high probability, and thus the method is far from practical.

Despite the objections we listed against the use of second-order methods, substantial effort has been devoted to making them efficient for DNNs, and exploiting curvature is still considered crucial. Trust-region methods [35] and cubic regularization methods [36, 37] based on second-order models are particularly well designed for the task. Moreover, these methods are now equipped with convergence rate analysis for nonconvex problems that include convergence to the first- and second-order stationary points. A comprehensive description of the worst-case global complexity in terms of convergence to first- and second-order stationary points for de-

[2] This is a form of automatic differentiation that uses the structure of the deep neural network.

[3] Because of nonconvexity we do not have global optimum guarantees.

terministic line search, trust-region, and adaptive cubic regularization methods can be found in [38, 39] and references therein.

Recently, some *randomized* methods for nonconvex optimization have been introduced. In particular, a nonconvex variant of SVRG has been proposed and analyzed in [40]. Convergence rates of a generic class of optimization methods that include line search and adaptive cubic regularization but use random models of the objective[4] have been established [41]. In particular, the convergence rate of such a randomized method has been shown to be the same as that of its fully deterministic counterpart, save for the constant multiple of $O(\frac{1}{2p-1})$, where p is the fixed probability that the random models provide sufficiently good approximation of the objective function. Unlike SVRG and other recent randomized methods, the methods in [41] do not assume any special form of the objective function; however, it is assumed that the function values (but not the derivatives) can be computed accurately. The reason this assumption is made is to allow the methods to employ adaptive step sizes, which tend to significantly improve the performance of optimization methods (especially in the nonconvex setting). SG methods, on the other hand, do not use adaptive step sizes and are sensitive to the step-size selection.

Fully stochastic variance-reducing trust-region methods have been developed during the past couple of years [42–44] and use adaptive trust-region size in the same way as their deterministic counterparts. Unlike the randomized methods, they do not exploit any particular form of the objective function or assume that its values can be computed accurately. They do assume that the function values (and possibly derivatives) can be approximated sufficiently well with high probability. Convergence to a first-order stationary point has been shown for these methods. Moreover, the method, assumptions, and convergence analysis in [42] led to a recent convergence rate result for a stochastic trust-region method for nonconvex optimization [45]. The convergence rate is the same as that of a deterministic method in the following sense: The expected number of iteration it takes the method to reach an iterate X^i for which $\|\nabla f(X^i)\|^2 \le \epsilon$ is bounded by $O(1/\epsilon)$. This result shows that a relatively standard trust-region methodology can be efficient in the stochastic setting and may outperform stochastic gradient descent.

In conclusion, trust-region methods may play a special role in optimizing deep neural networks. However, implementing them efficiently in that setting is an ongoing effort.

REFERENCES

[1] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[2] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. Technical Report 1606.04838, arXiv, 2016.

[3] J.R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, New York, 2011.

[4] R. Pasupathy and S. Ghosh. Simulation optimization: A concise overview and implementation guide. In *TutORials in Operations Research*, chapter 7, pages 122–150. INFORMS, 2013.

[5] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951.

[6] A. Ruszczynski and A. Shapiro, editors. *Stochastic Programming*. Handbooks in Operations Research and Management Science, Volume 10. Elsevier, Amsterdam, 2003.

[7] V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[8] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence,. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

[9] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, 2002.

[10] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, March 2003.

[11] G. Gnecco and M. Sanguineti. Approximation error bounds via Rademacher's complexity. *Applied Mathematical Sciences*, 2(4):153–176, 2008.

[12] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

[13] Si S., C.-J. Hsieh, and I. S. Dhillon. Memory efficient kernel approximation. In *International Conference on Machine Learning (ICML)*, pages 701–709, 2014.

[14] Mangasarian O.L. and Musicant D.R. Active set support vector machine classification. In T.K. Leen, V. Tresp, and T. G. Dietterich, editors, *Advances in Neural Information Processing Systems 13*, pages 577–583. MIT Press, Cambridge, MA, 2000.

[15] K. Scheinberg. An efficient implementation of an active set method for SVM. *Journal of Machine Learning Research*, pages 2237–2257, 2006.

[16] J. Gondzio. Interior point methods in machine learning. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*. MIT Press,, 2010.

[17] G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(1):335–366, January 2014.

[18] B. Recht, M. Hardt, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[19] O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 161–168. Curran Associates, Inc., 2008.

[20] R. Pasupathy, P. W. Glynn, S. Ghosh, and F. Hahemi. How much to sample in simulation-based stochastic recursions? Technical report, 2014. Under Review.

[21] R. Byrd, G. M. Chin, J. Nocedal, and Y. Wu. Sample size selection in optimization methods for machine learning. *Math. Program.*, 134(1):127–155, 2012.

[22] M. Schmidt, N. LeRoux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388*, 2013.

[23] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1646–1654. Curran Associates, Inc., 2014.

[24] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Math. Program.*, pages 1–41, 2013.

[4] For example, Taylor models based on a sample gradient with sufficiently large sample size.

[25] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. 2013.

[26] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Neurocomputing: Foundations of research. chapter Learning Representations by Back-Propagating Errors, pages 696–699. MIT Press, Cambridge, MA, 1988.

[29] B. A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994.

[30] W. Zhou, J. D. Griffin, and I. G. Akrotirianakis. A globally convergent modified conjugate-gradient line-search algorithm with inertia controlling. Technical Report 2009-01, 2009.

[31] J. Martens. Deep learning via Hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 735–742, 2010.

[32] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 2933–2941, Cambridge, MA, 2014. MIT Press.

[33] R. Pascanu, Y. N. Dauphin, S. Ganguli, and Y. Bengio. On the saddle point problem for non-convex optimization. *CoRR*, abs/1405.4604, 2014.

[34] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[35] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust Region Methods*. MPS/SIAM Series on Optimization. SIAM, Philadelphia, 2000.

[36] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Math. Program.*, 127(2):245–295, 2011.

[37] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Math. Program.*, 130(2):295–319, 2011.

[38] C. Cartis, N. I. M. Gould, and P. L. Toint. How much patience do you have? A worst-case perspective on smooth nonconvex optimization. *Optima*, 88, 2012.

[39] Y. Nesterov. *Introductory Lectures on Convex Optimization.* Kluwer Academic Publishers, Boston, MA, 2004.

[40] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. Technical Report arXiv:1603.06160, 2016.

[41] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. Technical Report, ISE, Lehigh, 2015.

[42] C. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. Technical Report arXiv:1504.04231, 2015.

[43] J. Larson and S.C. Billups. Stochastic derivative-free optimization using a trust region framework. *Computational Optimization and Applications*, 64(3):619–645, 2016.

[44] S. Shashaani, F. S. Hashemi, and R. Pasupathy. ASTRO-DF: a class of adaptive sampling trust-region algorithms for derivative-free simulation optimization. Technical report, Purdue University, 2015.

[45] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence rate analysis of a stochastic trust region method for nonconvex optimization. Technical Report arXiv:1609.07428, 2016.

In Memoriam

Jon Borwein: a personal reflection

Borwein at a 2006 summer school in Paseky.

Jon Borwein died on August 2, 2016. His untimely passing has deprived us all of a singular and brilliant mind and an inspirational intellectual leader, and I have lost a close personal friend. Rather than a formal memorial, my words are a personal reflection on my coauthor (of fifteen papers and a book [46]), a mentor to whom I owe my career.

Jon's mathematical breadth and energy make a fascinating but bewildering picture, extending far beyond traditional optimization, and challenging to sketch. He delighted in collaboration, and many of us knew first-hand his research style: whirling, exuberant, defamiliarizing, endlessly curious, elegant, scholarly, generous, and honest. He made time for everyone, no matter their rank or eccentricity. Shortly after I met Jon, at the height of the public prominence of his work around pi with his brother Peter, highlighted in their book [47] and a Scientific American article [48], he remarked to me how he kept in mind the eminent English mathematician G.H. Hardy, the sole reader of Ramanujan's first terse but prophetic notes.

Early in 1987 Jon had welcomed me to the delightful city of Halifax, Nova Scotia—then his home. During a two-year postdoctoral fellowship, I shadowed him closely on his travels. Astonishingly, among his many projects then was a Dictionary of Mathematics [49], and indeed I felt a kind of prosaic Boswell to his dizzying Samuel Johnson. In the decade that followed, we made our independent ways across Canada, through the University of Waterloo to Simon Fraser University. There, Jon founded the Center for Experimental and Computational Mathematics, a pioneering base for his international pre-eminence in experimental mathematics.

Jon laid down many roots. Wikipedia describes him as a "Scottish mathematician," born in St Andrews in 1951.

With his encyclopedic erudition, Jon probably knew Johnson's poke that "Much may be made of a Scotchman, if he be caught young"; if he did know, he took it to heart, receiving his doctorate as a Rhodes Scholar at Oxford. He went on to spend the core of his career in Canada, where he served as President of the Canadian Mathematical Society and was elected a Fellow of the Royal Society of Canada. Along the way, he was elected a Foreign Member of the Bulgarian Academy of Science, and Fellows of both the American Association for the Advancement of Science and the American Mathematical Society. He made his final home in 2009 at the University of Newcastle in Australia as Laureate Professor and was elected a Fellow of the Australian Academy of Science.

Jon's diverse honors make his generous and articulate collaborative style all the more striking. He worked and collaborated intensely but did not suffer fools gladly. I imagine he sympathized with another of Johnson's quips: "Sir, I have found you an argument; but I am not obliged to find you an understanding." He was nonetheless a painstaking and articulate stylist and in 1993 won (with his brother Peter and his long-time collaborator David Bailey) the mathematical world's highest honor for exposition, the Chauvenet Prize. (Sixty years earlier, the winner was G.H. Hardy.)

Jon and his family—his wife, Judi, and two young daughters, Rachel and Naomi (their sister Tova being yet to arrive)—more or less adopted me as a family member when I arrived in Canada. I essentially lived with them during month-long visits to Limoges, France (where Jon later received an honorary doctorate), to the Technion in Israel, and to Canberra and Newcastle, Australia. The sheer fun of that last visit probably inspired the Borweins' later choice of adopted home.

Life at the Borweins' home was an inspiring and exhausting blur. A typical evening involved prodigious and virtuoso culinary feats from Judi, feisty debates from Rachel and Naomi, and multiple simultaneous media playing at full volume. At a minimum, these included political news (Jon was intensely active, politically, serving for a while as treasurer of the Nova Scotia New Democratic Party), major league baseball (another domain of erudition), and music. All gradually dissolved into large glasses of Scotch (Jon's Scotchness, like Healey Willan's, was mostly "by absorbtion"), and then a call to arms from Jon to prove some reluctant theorem. The exuberant and dizzying environment mirrored Jon's mathematics, a style so appealing it quickly sealed my own career choice as a mathematician.

Jon left us too soon. I seek some small solace in that during any of his years, Jon's ideas were at least twice as good, twice as fast, twice as many, and twice as well shared as, say, mine. But for his beloved and devoted family, his death has been simply shocking and untimely.

Optimization theory was just one of Jon's arenas; but as the one I know best, I would like to pick out a few personal favorites, most from that same era. To Jon's extraordinary academic family of collaborators, many of whom I race by unmentioned, my apologies.

A theme running through much of Jon's work was his emphatic belief in optimization and analysis as a single discipline, often unified through the language of set-valued mappings. He recognized early, for example, the importance of characterizing "metric regularity" for constraint systems [50]–now commonly known as "error bounds," stably bounding the distance to the feasible region by a multiple of the constraint error. Such bounds are of widespread interest, in particular, in the convergence analysis of first-order methods. Jon and his student Heinz Bauschke used similar ideas in a deep and long-running study of von Neumann's alternating projection algorithm and its relatives [51]. Another theme underlying much of Jon's work on the interplay of analysis and optimization was his extensive use both of proximal analysis (a technique growing out of viscosity solutions of PDEs and optimal control) [52] and of generalized derivatives in surprising contexts, especially Banach space geometry [53].

Perhaps Jon's most celebrated result in nonsmooth analysis and optimization is the Borwein-Preiss variational principle [54]. A ubiquitous technique throughout variational mathematics appeals to the existence of a minimizer of a function. Without some compactness, the argument breaks, but a famous result of Ekeland rescues it through a small perturbation to the original function. Ekeland's perturbation is, unfortunately, nonsmooth; but using a deep and surprising argument, Borwein and Preiss showed that a smooth perturbation will in fact suffice.

Much of Jon's broader mathematical opus is intertwined with computation, and he believed fundamentally in the computer as a tool for mathematical discovery. Many of his contributions to optimization were, by contrast, conceptual rather than computational. An interesting exception is the Barzilai-Borwein method [55], an odd and ingenious nonmonotone gradient-based minimization algorithm that has attracted growing attention during the big-data-driven resurgence of first-order methods.

I cannot resist a nod at my own postdoctoral work with Jon, much of which grew out of the maximum entropy methodology for estimating an unknown probability density from some of its moments. In one of my favorite results from that period, we showed that a sequence of densities, the kth of which agreeing with the unknown density up to the first k moments, need not converge weakly in the space L1, but nonetheless must do so if each has the maximum possible Shannon entropy [56, 57].

Jon somehow maintained a fresh, youthful intellectual style until the end. Sitting on my desktop, dated two weeks before he died, is his last paper [58], a lovely essay on the craft of mathematical research. He writes: "I can no longer resist making some observations... to a young mathematician... but we are nearly all young longer." Fortunately, he shared his advice in the nick of time. His final instruction is "Above all, be honest."

The book [46] that Jon and I published together in 2000 has found some popularity, even though we had intended the material just to be a quick introduction, Chapter 1 of a serious book. It exists only because I momentarily caught up: for a brief and happy time, I could scribe slightly faster than Jon's genius could create. The subsequent chapters will not be done before we meet again.

Adrian Lewis
School of Operations Research and Information Engineering, Cornell University, USA, adrian.lewis@cornell.edu

REFERENCES

[46] J.M. Borwein and A.S. Lewis, *Convex Analysis and Nonlinear Optimization*, Springer, New York (2000).

[47] J.M. Borwein and P.B. Borwein, *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity*, Wiley, New York (1987).

[48] J.M. Borwein and P.B. Borwein, "Ramanujan and Pi," *Scientific American* (February 1988), 112–117.

[49] E.J. Borowski and J.M. Borwein, *Dictionary of Mathematics*, Collins, Glasgow (1989).

[50] J.M. Borwein, "Stability and regular points of inequality systems," *J. Optimization Theory and Applications* 48 (1986), 9–52.

[51] H.H. Bauschke and J.M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM Review* 38 (1996), 367–426.

[52] J.M. Borwein and A. Ioffe, "Proximal analysis in smooth spaces," *Set-Valued Analysis* 4 (1996), 1–24.

[53] J.M. Borwein, S.P. Fitzpatrick, and J.R. Giles, "The differentiability of real functions on normed linear space using generalized gradients," *J. Optimization Theory and Applications* 128 (1987), 512–534.

[54] J.M. Borwein and D. Preiss, "A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions," *AMS Transactions* 303 (1987), 517–527.

[55] J. Barzilai and J.M. Borwein, "Two point step-size methods," *IMA J. Numerical Analysis* 8 (1988), 141–148.

[56] J.M. Borwein and A.S. Lewis, "Convergence of best entropy estimates," *SIAM J. Optimization* 1 (1991), 191–205.

[57] J.M. Borwein and A.S. Lewis, "On the convergence of moment problems," *AMS Transactions* 325 (1991), 249–271.

[58] J.M. Borwein, "Generalisations, examples and counter-examples in analysis and optimisation," *Set-Valued and Variational Analysis* (2016), in press. DOI:10.1007/s11228-016-0379-2

Roger Fletcher

Roger Fletcher in his beloved highlands.

Roger Fletcher was one of the giants of optimization. He won the Dantzig Prize in 1997 and the Lagrange Prize in 2006, and he was a fellow of the Royal Society of Edinburgh, the Royal Society of London, and SIAM. Yet despite all these prizes and accolades, Roger remained humble and approachable throughout his career and never lost the sheer joy of solving problems and finding things out (he described his discovery of filter methods as making him "tingle when [he] thought about it").

Roger started his optimization at the dawn of nonlinear programming. Having completed his undergraduate degree at Cambridge on a state scholarship, he went to Leeds University to study for a Ph.D. At the time, Leeds had just set up one of the first computing labs in the UK. At Leeds, Roger first met Mike Powell and started a life-long competitive friendship. Mike had been invited to Leeds to present a seminar on derivative-free optimization but changed his mind at the last minute and asked whether he could instead talk about a recent technical report from Argonne by someone called Bill Davidon. As it happened, Roger had been given that same report by his advisor, Colin Reed. When Mike gave his seminar, he found that Roger already knew about this new method by Davidon and had a working code. So the two joined forces to write Roger's first paper, which appeared in the *Computer Journal* in 1963 [59], on what was later to be known as the DFP method.

The DFP paper was not the only hugely influential paper to come out of Roger's time at Leeds. Following a suggestion by his advisor, Roger investigated the use of his line search to extend the conjugate gradient method to nonlinear functions. The result was the Fletcher-Reeves method [60]. While neither of these two methods is much in use today, they represented a transformational moment in nonlinear programming. Roger modestly attributed the credit for these two first papers to his coauthors: "I had two huge ideas given to me by people." He described his time at Leeds as "probably the happiest period of [my] life". Many years later, while a professor at Dundee, he would create a similarly happy environment for his students.

Roger and Mike remained competitive friends throughout their careers. Roger once recounted a story about when he and Mike went hiking in the Scottish highlands. As Roger was slightly younger, he was leaving Mike behind, a situation that did not sit well with Mike's competitive nature. So Mike cleverly asked, "Roger, tell me about the proof of the conjugate gradient method"—and deftly managed to catch up with an out-of-breath Roger.

Roger always placed great importance on applications to drive his research. He believed that applications and numerical work help us understand the challenges of problems and the limitations of our algorithms. At Leeds, Roger was working on molecular dynamics calculations that required the solution of an unconstrained optimization problem, and his work on the DFP and conjugate-gradient methods was directly motivated by these applications. Later, Roger worked closely with the School of Chemical Engineering at the University of Edinburgh. He believed that applied mathematics research should ultimately be driven by what problems people wanted to solve. Roger also believed in small examples to show limitations of methods. Recently, he and Yuhong Dai developed a small example that showed that the Barzilai-Borwein method can fail for box-constrained problems [63].

A second foundation of Roger's research philosophy was software. He believed that software validates theory and at the same time is a guide to good methods. He wrote a powerful and robust LP solver and later `bqpd` [66], a Fortran77 package for solving indefinite QPs under degeneracy. These solvers supported both single- and double-precision arithmetic, because numerical difficulties would manifest themselves first in a single-precision version. His solver shows some ingenious object-orientesque coding in Fortran77 (even though I am sure Roger never knew anything about object-oriented programming or C++)! The QP solver relies on a matrix algebra "class" that implements the factorization of the basis matrix. Roger provided both dense and sparse instantiations of this "class" and opened the possibility for other classes, for example, for people wishing to exploit the structure of their problem.

Throughout his career, Roger distrusted textbooks. While he was working for his Ph.D., he implemented the steepest descent method, which was the method of choice at the time. It failed to solve his problem, and he grew distrustful of things written in books:

> I read in a couple of books (Householder's book [61], I think it was; another book, Hildebrand [62] perhaps). They seemed to suggest that steepest descent was a very good method. So I tried it, and it generated reams and reams of paper, punched paper tape output as the iterations progressed, and didn't make a lot of progress.[1]

However, Roger was just as suspicious of his own opinion, and not above changing his own mind. When he refereed a paper by a young Chinese mathematician on the Barzilai-Borwein method, he initially rejected the idea as useless. Luckily, the young Chinese mathematician persisted; and the result was that Roger not only changed his mind but also coauthored a number of papers with Yuhong Dai [63, 65, 64].

It is this self-doubt and suspicion that enabled Roger to stay fresh and produce groundbreaking results even late in his career. When I last met him, he was excited about his recent work on an augmented Lagrangian method for nonnegative QP. He was using a clever transformation of variables that also allowed him to store a single vector of size n that combines the primal and dual variables, thereby exploiting their natural complementarity.

Roger had a great sense of humor (if a somewhat limited reservoir of jokes). His deadpan humor was famous. On one occasion, he complemented me, "Nice pullover, is it new?", which confused me given Roger's lack of fashion sense and the fact that the pullover was quite old. The mystery was solved when I took the pullover off and discovered a gaping hole in its sleeve.

The author and Roger Fletcher at the summit.

Roger was what Americans would call a no-nonsense applied mathematician who believed in simple arguments and proofs. At the University of Dundee he fostered a "happy environment" for his many Ph.D. students and visitors. He selflessly provided guidance to his students, passing to a new generation of researchers the luck and good ideas that he felt he was given. Roger passed away in June 2016 while hiking in his beloved Scottish highlands.

Sven Leyffer
Mathematics and Computer Science Division, Argonne National Laboratory, USA, leyffer@anl.gov

REFERENCES

[59] R. Fletcher and M.J.D. Powell, "A rapidly convergent descent method for minimization," *The Computer Journal* 6(2):163–168, 1963.

[60] R. Fletcher and C.M. Reeves, "Function minimization by conjugate gradients," *The Computer Journal* 7(2):149–154, 1964.

[61] A.S. Householder, *Principles of Numerical Analysis*, McGraw-Hill Book Company, New York (1953).

[62] F.B. Hildebrand, *Introduction to Numerical Analysis*, McGraw-Hill, New York (1956).

[63] Y.H. Dai and R. Fletcher, "Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming," *Numerische Mathematik* 100(1):21–47, 2005.

[64] Y.H. Dai and R. Fletcher, "New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds," *Mathematical Programming* 106(3):403–421, 2006.

[65] Y.H. Dai and R. Fletcher, "On the asymptotic behaviour of some new gradient methods," *Mathematical Programming* 103(3):541–559, 2005.

[66] R. Fletcher, "Resolving degeneracy in quadratic programming," *Annals of Operations Research* 46(2):307–334, 1993.

[1] S. Leyffer, "An Interview with Roger Fletcher," *Optima* 99, 2015.

Christodoulos Achilleus Floudas

My father often repeated in my childhood that choosing an advisor is the most important decision a graduate student makes. Whatever mistakes I made as a Ph.D. student, I did one thing exceptionally well: I was privileged to work for Professor Christodoulos Achilleus Floudas. Chris's commitment to excellence was inspiring; his belief in his students was motivating; his enthusiasm for global optimization was contagious.

Chris passed away of an apparent heart attack on August 14, 2016, while vacationing in Greece with his family. He is survived by his wife, Fotini, and their daughter, Ismini. Born August 31, 1959 in Ioannina, Greece, Chris earned a diploma of chemical engineering at the Aristotle University of Thessaloniki in 1982 and a Ph.D. in chemical engineering from Carnegie Mellon University in 1986. He subsequently moved to Princeton University where he remained for 29 years and, in 2007, was appointed Stephen C. Macaleer '63 Professor in Engineering and Applied Science. In 2015, Chris moved to Texas A&M University, where he was appointed director of the Texas A&M Energy Institute and Erle Nye '59 Chair Professor for Engineering Excellence at the Artie McFerrin Department of Chemical Engineering.

To name only a few of his accolades, Chris was a member of the National Academy of Engineering (2011), a SIAM fellow (2013), a Thompson Reuters Highly Cited Researcher in two consecutive years (2014, 2015), and a corresponding member of the Academy of Athens (2015). Chris was also an outstanding mentor and teacher. In 2007, he became the first recipient of Princeton University's (now annual) Graduate Mentoring Award; recognition for his teaching includes the Princeton University Engineering Council Teaching Award (1995) and the Aspen Tech Excellence in Teaching Award (1999). He supervised 34 Ph.D. students and 20 postdoctoral research associates; many of these are now internationally leading researchers or professors. Ten further Ph.D. students are working on their degrees.[1]

This reflection focuses on the contributions Chris made to mathematical optimization, with a particular emphasis on deterministic global optimization. I will mostly neglect his contributions to multiscale systems engineering, chemical process synthesis and design, process operations, computational chemistry, and molecular biology. My central thesis is that he fundamentally opened the field of deterministic global optimization, making the discipline industrially relevant and blazing a path where subsequent research could follow. His unforgettable contribution is to drive mathematical optimization toward everyday relevance: He pushed the tractability boundary of optimization theory, algorithms, and implementations using a multitude of applications.

Chris's Ph.D. thesis, supervised by Professor Ignacio Grossmann at CMU, investigated the automatic synthesis of heat exchanger networks [67]. Designing heat exchanger networks is a mixed-integer nonlinear program (MINLP) with long-reaching implications for energy efficiency; a quarter of the EU 2012 energy consumption came from industry, and industry uses 73% of this energy on heating and cooling [68]. Chris's contribution was to derive many possible superstructures from a mixed-integer linear programming approximation of the MINLP, fix the binary decision variables, and locally solve the resulting nonlinear program. Chris also used flexibility analysis to design networks that can handle uncertain flowrates and temperatures. Flexibility analysis is a max-min-max optimization problem quantifying how far parameters can deviate from their nominal values while maintaining operational feasibility [69]; the closest analogue in modern-day mathematical programming research is robust optimization [70].

Chris developed a passion for deterministic global optimization soon after he started as an assistant professor in Princeton University's Department of Chemical Engineering in 1986. He and his early students developed the global optimization algorithm GOP, an extension of generalized Benders decomposition applicable to biconvex optimization problems [71–73]. Rigorous global optimization was not new, but Chris was the first to develop global optimization methodology and then apply these algorithms to important classes of nonconvex nonlinear chemical engineering problems, including the pooling problem [74] and phase equilibrium [75].

Chris's interest in computational chemistry led him to develop αBB, a convexification methodology based on the difference of convex functions. Energy minimization plays a central role in understanding and predicting the behavior of natural systems. Using an application to Lennard-Jones microclusters [76], Chris and Costas Maranas became the first to apply rigorous global optimization to molecular thermodynamics; energy minimization using local rather than global optimization may yield significant qualitative and quantitative errors in engineering decision making [77]. To find the global minimum energy configuration of Lennard-Jones microclusters, Maranas and Floudas [76] transformed the optimization problem, with its nonconvex pair potential term $(1/r^{12} - 2/r^6)$, into a difference-of-convex program using a sufficiently large parameter $\alpha > 0$ multiplied by a quadratic function. In autumn 1992, Chris took a sabbatical at Imperial College London where, with Wenbin Liu, he generalized the Maranas and Floudas [76] result to any nonconvex function with a Lipschitz continuous gradient [78]. These results

[1] Professor Floudas' academic tree: http://titan.engr.tamu.edu/tree/caf/

led to the now-famous αBB (α-branch-and-bound) convexification methodology where a convex lower bounding function $g^{\text{cnv}}(\mathbf{x})$ can be defined for generic nonconvex functions $g(\mathbf{x})$ with the addition of a separable convex quadratic function [79]:

$$g^{\text{cnv}}(\mathbf{x}) = g(\mathbf{x}) - \sum_i \alpha_i (x_i^U - x_i)(x_i - x_i^L),$$
$$\text{where } \alpha_i \geq \max\left\{0, -\frac{1}{2} \min_{\mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^U} \lambda(\mathbf{x})\right\}.$$

Chris and his coworkers developed protocols to automatically calculate the α parameters and implemented αBB-based global optimization as a generic MINLP solver [79–81]. Subsequent advances in αBB theory identified alternative univariate exponential perturbations [82], piecewise underestimators [83, 84], and nondiagonal αBB [85, 86]. The dependence on the eigenvalues λ of the Hessian matrix of the second derivatives of $g(\mathbf{x})$ has inspired further investigation into Hessian matrix spectral bounds [87].

The MINLP solver implementing the αBB methodology, written with Claire Adjiman and Yannis Androulakis [88], became Chris's major vehicle to approach a wide variety of applications: parameter estimation [89, 90], optimal control [91], reactor network synthesis [92], semi-infinite programming [93], bilevel programming [94], molecular structure prediction [95], peptide and protein folding [96–101], design under uncertainty [102], constrained systems of equations [103], phase equilibrium calculations [75, 104–106], and solution of subproblems within an augmented Lagrangian framework [107]. In many of these domains, Chris was the first researcher to even attempt applying global optimization.

But Chris was never interested in toy problems or proofs of concept. If he was going to apply mathematical optimization to a new problem class, he would always push optimization to its limit via highly relevant, industrially motivated test cases. For example, the problem of *de novo* peptide design begins with a flexible, 3D protein structure and finds amino acid sequences folding into the given template; the possible impact on both fundamentally understanding protein structure and developing novel therapeutics is enormous [108, 109]. The first stage of Chris's *de novo* peptide design framework is to develop many possible amino acid sequences; Chris and his coworkers identified a deep connection with the quadratic assignment problem (QAP) and adapted QAP solution techniques to the specific problem [110, 111]. In the second stage, Chris and his coworkers applied their ASTRO-FOLD technology (which uses αBB) to generate a 3D peptide structure [96–101]. This framework for *de novo* peptide design has been applied to several therapeutic applications including cancer [112] and HIV [113, 114].

Energy systems engineering is another domain that inspired Chris to make fundamental mathematical optimization contributions. As with molecular biology, Chris's complete refusal to address anything but the most pressing real-world problems led him to the tractability boundary of mathematical optimization. Together with Rich Baliban and Josephine Elia, Chris designed an optimization framework for the simultaneous process synthesis, heat and power integration of a thermochemical hybrid biomass, coal, and natural gas facility [115]. I remember raising some objections when Chris said that now we had to globally optimize a model with 15,439 continuous variables, 30 binary variables, 15,406 equality constraints, 230 inequality constraints, and 335 nonconvex terms [116]. But Chris absolutely insisted that his piecewise linear underestimation technique for bilinear terms [117–119] was not relevant unless it was effectively applied to real-world instances. Subsequently, Chris and his coworkers applied a similar methodology to pressing applications such as designing biomass-based production of chemicals and fuels [120–122] and transforming municipal solid waste to liquid fuel [123].

The difficulty of the aforementioned energy systems applications is primarily rooted in the nonconvex bilinear terms (e.g., $f \cdot c$), that arise with intermediate storage in process networks [124]. The special structure of the so-called pooling problem is fascinating: Flow variables f are multiplied by concentration variables c, and an undirected graph representation of the bilinear terms (where the decision variables are the nodes and the connecting edges are weighted by the coefficients) is bipartite [125]. Chris had made major contributions to this problem since 1990 [74], but one of his more recent intuitions was that piecewise linear underestimators seemed to be effective in solving the pooling problem [117–119]. Subsequent results further unpacked this relationship between the NLP pooling problem and MIP approximations [126], but still Chris would push me: Why did these piecewise approximations work so well? How could we eliminate the computational dependence on parameters? Actually, thanks to work with Radu Baltean-Lugojan [127], I finally have a first answer to these questions Chris asked me eight years ago; I had been so looking forward to upcoming conferences to tell him all about it.

Chris's incorrigible, unquenchable enthusiasm for global optimization was contagious. When I was working with Chris, he had already developed, with Cliff Meyer, several important results on the facets of the trilinear monomials [128, 129]. Chris and Cliff had further generalized these results to automatically generate the facets of *edge-concave functions*, for example, those admitting a vertex polyhedral envelope [130]. One of my projects extended the Meyer and Floudas [130] methodology to develop multiterm underestimators for bilinear functions [125]. One day, when Chris was on a business trip, I sent him an email titled *good news (regarding facets of bilinear functions)* and received this reply on March 12, 2011 at 08:28 EST:

```
Dear Ruth,
This is great!!! I would like to see what you and
Jamie have upon my return. I have some additional
ideas that I would like to discuss too.
With best regards,
Chris
```

I learned later that Chris had been at the Narita airport in Tokyo during the 2011 Tōhoku earthquake that hit 14:46 JST on March 11. Chris was himself a force of nature: Just 24 hours after the aftershocks from the fourth most powerful earthquake ever recorded had mostly stopped, Chris was back to actively encouraging his Ph.D. students in their projects.

Chris's belief in his students was inspiring. In 2008, when Chris told me that my early-stage Ph.D. assessment needed a promise to write and publicly release a global MINLP solver, I told him he was being unreasonable. He said to do it anyway; and I must have made a face, because I remember him telling me sharply: "you will not roll your eyes at your advisor!" He was right in the end, and we released first GloMIQO [131] and then ANTIGONE [132] as global optimization solvers in the modeling language GAMS. I really enjoyed developing the solvers, which incorporated several contributions Chris and his coworkers made, including those to αBB [80], generalised geometric programming [133, 134], edge-concave underestimators [130, 135], and automatic transformations for process networks problems [125, 131].

But, more than anything, I appreciated how Chris made my work relevant by insisting that it could apply to real-world problems. One application of ANTIGONE (although I had nothing to do with it), involved Chris's developing, with Fani Boukouvala and Faruque Hasan, a general methodology for grey-box models [136]. Chris had previously developed approaches in black-box optimization [137], optimization for functions without explicit analytical representation, but his new work was especially exciting because it develops optimization methodology for applications where some of the governing equations are known but others have to be determined using a data-driven methodology [138]. Chris and his coworkers used this methodology to develop a global optimization framework for entire petrochemical planning operations [139].

Before I close, I need to mention some of the contributions that Chris made to optimization under uncertainty. Chris first encountered uncertainty in synthesizing heat exchanger networks with uncertain flowrates and temperatures [69], and uncertainty was thereafter a thread that ran throughout his research [140]. Chris's early work in optimization under uncertainty was mostly in flexibility [102], but he later became keenly interested in robust optimization and introduced the methodology to the chemical engineering community [141–144]. In particular, Chris and his coworkers extended robust optimization, which was originally developed as a convex optimization methodology, to important nonconvex problems including operational planning [145], crude oil scheduling [146], and the vehicle-routing problem [147].

The contributions Chris made to mathematical optimization are legion, but his contributions also extended to supporting the work of many other global optimizers. Together with Panos Pardalos, Chris and his students published the *Handbook of Test Problems in Local and Global Optimization*, which has become a standard test set for any global optimization solver [148]. Chris and Pardalos edited many other special issues and books together [149–153]; the most significant is the *Encyclopedia of Optimization* [154]. These books brought significant attention to deterministic global optimization and supported the careers of many researchers. Also deserving mention is the close friendship between Chris and Stratos Pistikopoulos; the two met as undergraduates and forged a tight connection while both were Ph.D. students at CMU. They were collaborators [73, 140, 155, 156], but another lasting result of Chris and Stratos's friendship was their roles as founding director and associate director, respectively, of the Texas A&M Energy Institute. There the work of Chris continues: The Ph.D. students of Chris are all keen to continue with Stratos the work they started with Chris.

Chris Floudas and the author at the 2014 AIChE Annual Meeting.

Ruth Misener
Department of Computing, Imperial College London, UK, r.misener@imperial.ac.uk

REFERENCES

[67] Floudas, Ciric, and Grossmann. Automatic synthesis of optimum heat exchanger network configurations. *AIChE J*, 32(2):276–290, 1986.

[68] European Commission. An EU strategy on heating and cooling, 2016. Brussels, COM(2016) 51 final.

[69] Floudas and Grossmann. Synthesis of flexible heat-exchanger networks with uncertain flowrates and temperatures. *Comput Chem Eng*, 11(4):319–336, 1987.

[70] Zhang, Grossmann, and Lima. On the relation between flexibility analysis and robust optimization for linear systems. *AIChE J*, 62(9):3109–3123, 2016.

[71] Floudas and Visweswaran. A global optimization algorithm (GOP) for certain classes of nonconvex NLPs: I. Theory. *Comput Chem Eng*, 14(12):1397–1417, 1990.

[72] Floudas and Visweswaran. Primal-relaxed dual global optimization approach. *J Optim Theory Appl*, 78(2):187–225, 1993.

[73] Visweswaran, Floudas, Ierapetritou, and Pistikopoulos. A decomposition-based global optimization approach for solving bilevel linear and quadratic programs. In Floudas and Pardalos, eds., *State of the Art in Global Optimization: Computational Methods and Applications*, volume 7 of *Nonconvex Optimization and its Applications*, pages 139–162, 1996.

[74] Visweswaran and Floudas. A global optimization algorithm (GOP) for certain classes of nonconvex NLPs: II. application of theory and test problems. *Comput Chem Eng*, 14(12):1419–1434, 1990.

[75] McDonald and Floudas. Decomposition based and branch and bound global optimization approaches for the phase equilibrium problem. *J Glob Optim*, 5:205–251, 1994.

[76] Maranas and Floudas. A global optimization approach for Lennard-Jones microclusters. *J Chem Phys*, 97(10):7667–7678, 1992.

[77] Bollas, Barton, and Mitsos. Bilevel optimization formulation for parameter estimation in vapor-liquid(-liquid) phase equilibrium problems. *Chem Eng Sci*, 64(8):1768–1783, 2009.

[78] Liu and Floudas. A remark on the GOP algorithm for global optimization. *J Glob Optim*, 3:519–521, 1993.

[79] Androulakis, Maranas, and Floudas. αBB: A global optimization method for general constrained nonconvex problems. *J Glob Optim*, 7:337–363, 1995.

[80] Adjiman, Dallwig, Floudas, and Neumaier. A global optimization method, αBB, for general twice differentiable NLPs-I. Theoretical advances. *Comput Chem Eng*, 22:1137–1158, 1998.

[81] Adjiman, Androulakis, and Floudas. A global optimization method, αBB, for general twice differentiable NLPs-II. Implementation and computational results. *Comput Chem Eng*, 22:1159–1179, 1998.

[82] Akrotirianakis and Floudas. Computational experience with a new class of convex underestimators: Box-constrained NLP problems. *J Glob Optim*, 29(3):249–264, 2004.

[83] Meyer and Floudas. Convex underestimation of twice continuously differentiable functions by piecewise quadratic perturbation: Spline αBB underestimators. *J Glob Optim*, 32(2):221–258, 2005.

[84] Gounaris and Floudas. Tight convex underestimators for $\mathcal{C}^2$-continuous problems: II. Multivariate functions. *J Glob Optim*, 42(1):69–89, 2008.

[85] Akrotirianakis, Meyer, and Floudas. The role of the off-diagonal elements of the Hessian matrix in the construction of tight convex underestimators for nonconvex functions. In *Discovery Through Product and Process Design*, pages 501–504. Foundations of Computer-Aided Process Design, 2004.

[86] Skjäl, Westerlund, Misener, and Floudas. A generalization of the classical αBB convex underestimation via diagonal and nondiagonal quadratic terms. *J Optim Theory Appl*, 154(2):462–490, 2012.

[87] Mönnigmann. Efficient calculation of bounds on spectra of Hessian matrices. *SIAM J Sci Comput*, 30(5):2340–2357, 2008.

[88] Adjiman, Androulakis, and Floudas. Global optimization of mixed-integer nonlinear problems. *AIChE J*, 46(9):1769–1797, 2000.

[89] Esposito and Floudas. Global optimization in parameter estimation of nonlinear algebraic models via the error-in-variables approach. *Ind Eng Chem Res*, 37:1841–1858, 1998.

[90] Esposito and Floudas. Global optimization for the parameter estimation of differential-algebraic systems. *Ind Eng Chem Res*, 39(5):1291–1310, 2000.

[91] Esposito and Floudas. Deterministic global optimization in nonlinear optimal control problems. *J Glob Optim*, 17:97–126, 2000.

[92] Esposito and Floudas. Deterministic global optimization in isothermal reactor network synthesis. *J Glob Optim*, 22(1-4):59–95, 2002.

[93] Floudas and Stein. The adaptive convexification algorithm: A feasible point method for semi-infinite programming. *SIAM J Optim*, 18(4):1187–1208, 2007.

[94] Gümüs and Floudas. Global optimization of nonlinear bilevel programming problems. *J Glob Optim*, 20(1):1–31, 2001.

[95] Maranas and Floudas. A deterministic global optimization approach for molecular-structure determination. *J Chem Phys*, 100(2):1247–1261, 1994.

[96] Klepeis and Floudas. Free energy calculations for peptides via deterministic global optimization. *J Chem Physics*, 110:7491–7512, 1999.

[97] Klepeis and Floudas. Deterministic global optimization and torsion angle dynamics for molecular structure prediction. *Comput Chem Eng*, 24(2-7):1761–1766, 2000.

[98] Klepeis and Floudas. Ab initio tertiary structure prediction of proteins. *J Glob Optim*, 25:113–140, 2003.

[99] Klepeis and Floudas. ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical J*, 85:2119–2146, 2003.

[100] Klepeis, Floudas, Morikis, and Lambris. Predicting peptide structures using NMR data and deterministic global optimization. *J Comput Chem*, 20(13):1354–1370, 1999.

[101] Klepeis, Pieja, and Floudas. A new class of hybrid global optimization algorithms for peptide structure prediction: Integrated hybrids. *Comput Phys Commun*, 151:121–140, 2003.

[102] Floudas, Gümüs, and Ierapetritou. Global optimization in design under uncertainty: Feasibility test and flexibility index problems. *Ind Eng Chem Res*, 40(20):4267–4282, 2001.

[103] Maranas and Floudas. Finding all solutions of nonlinearly constrained systems of equations. *J Glob Optim*, 7(2):143–182, 1995.

[104] McDonald and Floudas. Global optimization for the phase and chemical equilibrium problem: Application to the NRTL equation. *Comput Chem Eng*, 19(11):1111–1141, 1995.

[105] McDonald and Floudas. Global optimization for the phase stability problem. *AIChE J*, 41(7):1798–1814, 1995.

[106] McDonald and Floudas. Global optimization and analysis for the Gibbs free energy function for the UNIFAC, Wilson, and ASOG equations. *Ind Eng Chem Res*, 34:1674–1687, 1995.

[107] Birgin, Floudas, and Martínez. Global minimization using an augmented Lagrangian method with variable lower-level constraints. *Math Program*, 125:139–162, 2010.

[108] Floudas, Fung, McAllister, Mönnigmann, and Rajgaria. Advances in protein structure prediction and de novo protein design: A review. *Chem Eng Sci*, 61(3):966–988, 2006.

[109] Khoury, Smadbeck, Kieslich, and Floudas. Protein folding and de novo protein design for biotechnological applications. *Trends in Biotechnology*, 32(2):99–109, 2014.

[110] Klepeis, Floudas, Morikis, Tsokos, and Lambris. Design of peptide analogues with improved activity using a novel de novo protein design approach. *Ind Eng Chem Res*, 43(14):3817–3826, 2004.

[111] Fung, Rao, Floudas, Prokopyev, Pardalos, and Rendl. Computational comparison studies of quadratic assignment like formulations for the in silico sequence selection problem in de novo protein design. *J Comb Optim*, 10(1):41–60, 2005.

[112] Smadbeck, Peterson, Zee, Garapaty, Mago, Lee, Giannis, Trojer, Garcia, and Floudas. De novo peptide design and experimental validation of histone methyltransferase inhibitors. *PLoS One*, 9(2):e90095, 2014.

[113] Bellows and Floudas. Computational methods for de novo protein design and its applications to the human immunodeficiency virus 1, purine nucleoside phosphorylase, ubiquitin specific protease 7, and histone demethylases. *Curr Drug Targets*, 11:264–278, 2010.

[114] Kieslich, Tamamis, Guzman, Onel, and Floudas. Highly accurate structure-based prediction of HIV-1 coreceptor usage suggests intermolecular interactions driving tropism. *PLoS ONE*, 11(2):1–15, 2016.

[115] Baliban, Elia, and Floudas. Optimization framework for the simultaneous process synthesis, heat and power integration of a thermochemical hybrid biomass, coal, and natural gas facility. *Comput Chem Eng*, 35(9):1647–1690, 2011.

[116] Baliban, Elia, Misener, and Floudas. Global optimization of a MINLP process synthesis model for thermochemical based conversion of hybrid coal, biomass, and natural gas to liquid fuels. *Comput Chem Eng*, 42:64–86, 2012.

[117] Meyer and Floudas. Global optimization of a combinatorially complex generalized pooling problem. *AIChE J*, 52(3):1027–1037, 2006.

[118] Gounaris, Misener, and Floudas. Computational comparison of piecewise-linear relaxations for pooling problems. *Ind Eng Chem Res*, 48(12):5742–5766, 2009.

[119] Misener, Thompson, and Floudas. APOGEE: Global optimization of standard, generalized, and extended pooling problems via linear and logarithmic partitioning schemes. *Comput Chem Eng*, 35(5):876–892, 2011.

[120] Niziolek, Onel, Elia, Baliban, Xiao, and Floudas. Coal and biomass to liquid transportation fuels: Process synthesis and global optimization strategies. *Ind Eng Chem Res*, 53(44):17002–17025, 2014.

[121] Matthews, Niziolek, Onel, Pinnaduwage, and Floudas. Biomass to liquid transportation fuels via biological and thermochemical conversion: Process synthesis and global optimization strategies. *Ind Eng Chem Res*, 55(12):3203–3225, 2016.

[122] Niziolek, Onel, Guzman, and Floudas. Biomass-based production of benzene, toluene, and xylenes via methanol: Process synthesis and deterministic global optimization. *Energy Fuels*, 30(6):4970–4998, 2016.

[123] Niziolek, Onel, and Floudas. Municipal solid waste to liquid transportation fuels, olefins, and aromatics: Process synthesis and deterministic global optimization. *Comput Chem Eng*, 2016. DOI: 10.1016/j.compchemeng.2016.07.024.

[124] Misener and Floudas. Advances for the pooling problem: Modeling, global optimization, and computational studies. *Appl Comput Math*, 8(1):3–22, 2009.

[125] Misener, Smadbeck, and Floudas. Dynamically-generated cutting planes for mixed-integer quadratically-constrained quadratic programs and their incorporation into GloMIQO 2.0. *Optim Met Softw*, 30(1):215–249, 2014.

[126] Dey and Gupte. Analysis of MILP techniques for the pooling problem. *Oper Res*, 63(2):412–427, 2015.

[127] Baltean-Lugojan and Misener. A parametric approach to the pooling problem, 2016. http://www.optimization-online.org/DB_HTML/2016/05/5457.html.

[128] Meyer and Floudas. Trilinear monomials with positive or negative domains: Facets of the convex and concave envelopes. In Floudas and Pardalos, eds., *Frontiers in Global Optimization*, pages 327–352. Kluwer, 2003.

[129] Meyer and Floudas. Trilinear monomials with mixed sign domains: Facets of the convex and concave envelopes. *J Glob Optim*, 29(2):125–155, 2004.

[130] Meyer and Floudas. Convex envelopes for edge-concave functions. *Math Program*, 103(2):207–224, 2005.

[131] Misener and Floudas. GloMIQO: Global Mixed-Integer Quadratic Optimizer. *J Glob Optim*, 57(1):3–50, 2013.

[132] Misener and Floudas. ANTIGONE: Algorithms for coNTinuous Integer Global Optimization of Nonlinear Equations. *J Glob Optim*, 59(2-3):503–526, 2014.

[133] Maranas and Floudas. Global optimization in generalized geometric programming. *Comput Chem Eng*, 21(4):351–369, 1997.

[134] Misener and Floudas. A framework for globally optimizing mixed-integer signomial programs. *J Optim Theory Appl*, 161:905–932, 2014.

[135] Misener and Floudas. Global optimization of mixed-integer quadratically-constrained quadratic programs (MIQCQP) through piecewise-linear and edge-concave relaxations. *Math Program*, 136:155–182, 2012.

[136] Boukouvala, Hasan, and Floudas. Global optimization of general constrained grey-box models: new method and its application to constrained PDEs for pressure swing adsorption. *J Glob Optim*, 2015. DOI: 10.1007/s10898-015-0376-2.

[137] Meyer, Floudas, and Neumaier. Global optimization with nonfactorable constraints. *Ind Eng Chem Res*, 41(25):6413–6424, 2002.

[138] Boukouvala and Floudas. Argonaut: Algorithms for global optimization of constrained grey-box computational problems. *Optim Lett*, 2016. DOI: 10.1007/s11590-016-1028-2.

[139] Li, Xiao, Boukouvala, Floudas, Zhao, Du, Su, and Liu. Data-driven mathematical modeling and global optimization framework for entire petrochemical planning operations. *AIChE J*, 62(9):3020–3040, 2016.

[140] Ierapetritou, Pistikopoulos, and Floudas. Operational planning under uncertainty. *Comput Chem Eng*, 20(12):1499–1516, 1996.

[141] Lin, Janak, and Floudas. A new robust optimization approach for scheduling under uncertainty: I. Bounded uncertainty. *Comput Chem Eng*, 28(6-7):1069–1085, 2004.

[142] Janak, Lin, and Floudas. A new robust optimization approach for scheduling under uncertainty: II. Uncertainty with known probability distribution. *Comput Chem Eng*, 31(3):171–195, 2007.

[143] Li and Floudas. Optimal scenario reduction framework based on distance of uncertainty distribution and output performance: I. Single reduction via mixed integer linear optimization. *Comput Chem Eng*, 70:50–66, 2014.

[144] Li and Floudas. Optimal scenario reduction framework based on distance of uncertainty distribution and output performance: II. Sequential reduction. *Comput Chem Eng*, 84:599–610, 2016.

[145] Verderame and Floudas. Operational planning of large-scale industrial batch plants under demand due date and amount uncertainty. I. Robust optimization framework. *Ind Eng Chem Res*, 48(15):7214–7231, 2009.

[146] Li, Misener, and Floudas. Scheduling of crude oil operations under demand uncertainty: A robust optimization framework coupled with global optimization. *AIChE J*, 58(8):2373–2396, 2012.

[147] Gounaris, Wiesemann, and Floudas. The robust capacitated vehicle routing problem under demand uncertainty. *Oper Res*, 61(3):677–693, 2013.

[148] Floudas, Pardalos, Adjiman, Esposito, Gümüs, Harding, Klepeis, Meyer, and Schweiger. *Handbook of Test Problems in Local and Global Optimization*. Kluwer, 1999.

[149] Floudas and Pardalos. State-of-the-art in global optimization - computational methods and applications - preface. *J Glob Optim*, 7(2):113, 1995.

[150] Floudas and Pardalos, eds. *State of the Art in Global Optimization: Computational Methods and Applications*. Kluwer, 1996.

[151] Floudas and Pardalos, eds. *Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches*. Nonconvex Optimization and Its Applications. Kluwer, 2000.

[152] Floudas and Pardalos, eds. *Frontiers in Global Optimization*. Nonconvex Optimization and Its Applications, Kluwer, 2004.

[153] Floudas and Pardalos, eds. *J Glob Optim*. 2009. *43*(2).

[154] Floudas and Pardalos. *Encyclopedia of optimization*. Springer Science & Business Media, 2001.

[155] Papalexandri, Pistikopoulos, and Floudas. Mass-exchange networks for waste minimization–a simultaneous approach. *Chem Eng Res Design*, 72(3):279–294, 1994.

[156] Floudas and Pistikopoulos. Professor Ignacio E. Grossmann-Tribute. *Comput Chem Eng*, 72:1–2, 2015.

Bulletin

Email items to siagoptnews@lists.mcs.anl.gov for consideration in the bulletin of forthcoming issues.

1 Event Announcements

1.1 SIAM Conference on Optimization (OP17)

SIAM Conference on
OPTIMIZATION
Based on Audet and Dennis' plot of Kolda, Lewis and Torczon's modification of the Dennis-Wood canoe function.
May 22-25, 2017
Sheraton Vancouver Wall Centre
Vancouver, British Columbia, Canada

The SIAM Conference on Optimization (OP17) will feature the latest research in theory, algorithms, software, and applications in optimization problems. A particular emphasis will be put on applications of optimization in health care, biology, finance, aeronautics, control, operations research, and other areas of science and engineering. The conference brings together mathematicians, operations researchers, computer scientists, engineers, software developers, and practitioners, thus providing an ideal environment to share new ideas and important problems among specialists and users of optimization in academia, government, and industry. Themes for this conference are as follows:

- Applications in Health Care
- Applications in Energy Networks and Renewable Resources Integer Optimization
- Applications in Machine Learning and Signal Processing
- Combinatorial and Mixed Integer Nonlinear Optimization
- Conic Optimization
- Derivative-free Optimization
- Graphs and Networks
- Nonlinear Optimization

Deadlines (Midnight Eastern Time)
October 24: Minisymposium Proposal Submissions
November 21: Contributed Lecture, Poster, and Minisymposium Presentation Abstracts

Plenary Speakers: Eva Lee, Georgia Institute of Technology, USA; Jeffrey Linderoth, University of Wisconsin-Madison, USA; Zhi-Quan (Tom) Luo, University of Minnesota, USA, and Chinese University of Hong Kong, Hong Kong; Ali Pinar, Sandia National Laboratories, USA; James Renegar, Cornell University, USA; Katya Scheinberg, Lehigh University, USA; Martin Wainwright, University of California at Berkeley, USA.

More details are available on the conference website: http://www.siam.org/meetings/op17/index.php.

1.2 SAMSI Workshop on the Interface of Statistics and Optimization (WISO)

This workshop, held in conjunction with SAMSI's Program on Optimization, will take place February 8–10, 2017.

Description: The integration and cross-fertilization between statistics and optimization is urgent and productive. Traditionally, optimization has merely been used as a tool to compute numerical solutions for statistical problems, while optimization theory and algorithms have rarely supported statistical techniques. This compartmental approach is proving to be non-optimal. More and more statistical tools are being developed by borrowing strengths from optimization, while optimization is looking to statistics for new insights, speed and robustness.

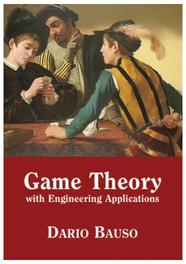
This workshop will bring together researchers who are pioneers in the synergy of statistics and optimization. It will serve SAMSIs mission to forge a synthesis of the statistical sciences and the applied mathematical sciences with disciplinary science to confront the very hardest and most important data- and model-driven scientific challenges. The workshop will also address contemporary issues with potentially significant impact in industrial applications.

Location: This workshop will be held at the Hamner Conference Center in Research Triangle Park.

More details are available on the conference website: https://www.samsi.info/programs-and-activities/research-workshops/workshop-interface-statistics-optimization-wiso/ and by contacting the organizers (Xiaoming Huo, Dirk Lorenz, Ekkehard Sachs, and Hua Zhou) and SAMSI directorate liaison (Ilse Ipsen).

2 Book Announcements

2.1 Game Theory with Engineering Applications



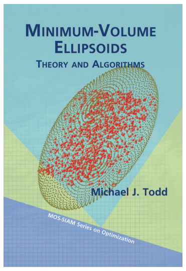
By Dario Bauso
Publisher: SIAM
Series: Advances in Design and Control, Vol. 30
ISBN: 978-1-611974-27-0, xxviii + 292 pages
Published: 2016
http://bookstore.siam.org/dc30/

About the book: Engineering systems are highly distributed collective systems–decisions, information, and objectives are distributed throughout–that have humans in the loop, and thus decisions may be influenced by socioeconomic factors. Engineering systems emphasize the potential of control and games beyond traditional applications. Game theory can be used to design incentives to obtain socially desirable behaviors on the part of the players, for example, a change in the consumption patterns of the "prosumers" (producers-consumers) or better redistribution of traffic.

This unique book addresses the foundations of game theory, with an emphasis on the physical intuition behind the concepts, an analysis of design techniques, and a discussion of new trends in the study of cooperation and competition in large complex distributed systems.

Audience: This book is intended for undergraduate and graduate students and researchers in industrial, aeronautical, manufacturing, civil, mechanical, chemical, and electrical engineering. It is also designed for social scientists interested in quantitative methods for sociotechnical systems, biologists working on adaptation mechanisms and evolutionary dynamics, and physicists working on collective systems and synchronization phenomena.

2.2 Minimum-Volume Ellipsoids: Theory and Algorithms



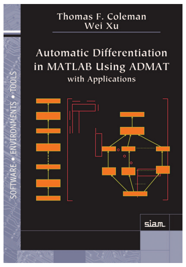
By Michael J. Todd
Publisher: SIAM
Series: MOS-SIAM Series on Optimization, Vol. 23
ISBN: 978-1-611974-37-9, xiv + 149 pages
Published: July 2016
http://bookstore.siam.org/MO23/

About the book: This book, the first on these topics, addresses the problem of finding an ellipsoid to represent a large set of points in high-dimensional space, which has applications in computational geometry, data representations, and optimal design in statistics. The book covers the formulation of this and related problems, theoretical properties of their optimal solutions, and algorithms for their solution. Due to the high dimensionality of these problems, first-order methods that require minimal computational work at each iteration are attractive. While algorithms of this kind have been discovered and rediscovered over the past fifty years, their computational complexities and convergence rates have only recently been investigated. The optimization problems in the book have the entries of a symmetric matrix as their variables, so the author's treatment also gives an introduction to recent work in matrix optimization.

This book provides historical perspective on the problems studied by optimizers, statisticians, and geometric functional analysts; demonstrates the huge computational savings possible by exploiting simple updates for the determinant and the inverse after a rank-one update, and highlights the difficulties in algorithms when related problems are studied that do not allow simple updates at each iteration; and gives rigorous analyses of the proposed algorithms, MATLAB codes, and computational results.

Audience: This book will be of interest to graduate students and researchers in operations research, theoretical statistics, data mining, complexity theory, computational geometry, and computational science.

2.3 Automatic Differentiation in MATLAB Using ADMAT with Applications


Thomas F. Coleman
Wei Xu
Automatic Differentiation in MATLAB Using ADMAT with Applications
SOFTWARE • ENVIRONMENTS • TOOLS
siam

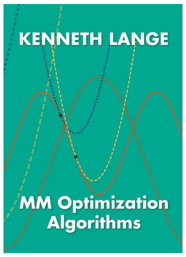
By Thomas F. Coleman & Wei Xu
Publisher: SIAM
Series: Software, Environments, and Tools, Vol. 27
ISBN: 978-1-611974-35-5, xii + 105 pages
Published: June 2016
http://bookstore.siam.org/SE27/

About the book: The calculation of partial derivatives is a fundamental need in scientific computing. Automatic differentiation (AD) can be applied straightforwardly to obtain all necessary partial derivatives (usually first and, possibly, second derivatives) regardless of a code's complexity. However, the space and time efficiency of AD can be dramatically improved—sometimes transforming a problem from intractable to highly feasible—if inherent problem structure is used to apply AD in a judicious manner.

Automatic Differentiation in MATLAB using ADMAT with Applications discusses the efficient use of AD to solve real problems, especially multidimensional zero-finding and optimization, in the MATLAB environment. This book is concerned with the determination of the first and second derivatives in the context of solving scientific computing problems with an emphasis on optimization and solutions to nonlinear systems. The authors focus on the application rather than the implementation of AD, solve real nonlinear problems with high performance by exploiting the problem structure in the application of AD, and provide many easy to understand applications, examples, and MATLAB templates.

Audience: This book will prove useful to financial engineers, quantitative analysts, and researchers working with inverse problems, as well as to engineers and applied scientists in other fields.

2.4 MM Optimization Algorithms



By Kenneth Lange
Publisher: SIAM
ISBN: 978-1-611974-39-3, x + 223 pages
Published: July 2016
http://bookstore.siam.org/OT147/

About the book: MM Optimization Algorithms offers an overview of the MM principle, a device for deriving optimization algorithms satisfying the ascent or descent property. These algorithms can separate the variables of a problem, avoid large matrix inversions, linearize a problem, restore symmetry, deal with equality and inequality constraints gracefully, and turn a nondifferentiable problem into a smooth problem.

The author presents the first extended treatment of MM algorithms, which are ideal for high-dimensional optimization problems in data mining, imaging, and genomics; derives numerous algorithms from a broad diversity of application areas, with a particular emphasis on statistics, biology, and data mining; and summarizes a large amount of literature that has not reached book form before.

Audience: This book is intended for those interested in high-dimensional optimization. Background material on convexity and semidifferentiable functions is derived in a setting congenial to graduate students.

3 Other Announcements

3.1 Howard Rosenbrock Prize 2015

The Howard Rosenbrock Prize is awarded each year for the best paper in *Optimization and Engineering (OPTE)*, a Springer journal that promotes the advancement of optimization methods and the innovative application of optimization in engineering. The prize is named after Howard

Rosenbrock, a pioneer of modern control theory and practice who embodied the bridging of the gap between optimization and engineering that is *OPTE*'s raison d'être.

The recipients for 2015 are Moritz Simon and Michael Ulbrich of the Department of Mathematics, Technische Universität München. Their awarded article is entitled "Adjoint based optimal control of partially miscible two-phase flow in porous media with applications to CO_2 sequestration in underground reservoirs." The formal announcement of the prize is published in the latest issue of *OPTE*.

Congratulations to Moritz and Michael!

Miguel Anjos, Editor-in-Chief, Optimization and Engineering, `https://www.optejournal.com`

3.2 Combinatorial Scientific Computing Best Paper Prize

The 7th SIAM Workshop on Combinatorial Scientific Computing (CSC16) was held October 10–12 in Albuquerque, New Mexico. A new feature of CSC16 was a peer-reviewed proceedings published by SIAM. The inaugural CSC Best Paper Prize was awarded to Fredrik Manne, Md. Naim, Hakon Lerring (all from the University of Bergen, Norway) and Mahantesh Halappanavar (Pacific Northwest National Laboratory) for their paper "Stable marriages and greedy matchings."

This paper showed that recent approximation algorithms for computing edge weighted matchings in graphs can be viewed as variants of the Gale-Shapley and Wilson-McVitie algorithms for the celebrated stable marriage problem. This paper links work performed by two distinct communities, leading to increased understanding and parallel algorithms.

This year's prize committee consisted of Uwe Naumann, Alex Pothen, and Sivan Toledo. Congratulations to Professor Manne and his coauthors.

3.3 2016 SIAM Fellows Announced

Each year, SIAM designates as Fellows of the society those who have made outstanding contributions to the fields of applied mathematics and computational science. This year, 30 members of the community were selected for this distinction.

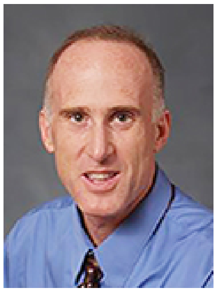
These new Fellows include eight members of the SIAG, whose citations are included below. Full details on the SIAM Fellow program can be found at `http://www.siam.org/prizes/fellows/index.php`. Congratulations to all the new Fellows!

Gang Bao
Zheijiang University
For significant and lasting contributions to inverse problems in wave phenomena and electromagnetics applied to optics.

Thomas F. Coleman
University of Waterloo
For contributions to large-scale, sparse numerical optimization, financial optimization, and leadership in mathematics education and engagement with industry.

Michael Hintermüller
Weierstrass Institute for Applied Analysis and Stochastics and Humboldt-Universität zu Berlin
For contributions to theoretical and numerical optimization, and for their application.

Andrew Knyazev
Mitsubishi Electric Research Laboratories (MERL) and Professor Emeritus at University of Colorado Denver
For contributions to computational mathematics and development of numerical methods for eigenvalue problems.



James G. Nagy
Emory University
For contributions to the computational science of image reconstruction.

Cynthia A. Phillips
Sandia National Laboratories
For contributions to the theory and applications of combinatorial optimization.

David P. Williamson
Cornell University
For fundamental contributions to the design and analysis of approximation algorithms.

Xunyu Zhou
Columbia University and University of Oxford
For accomplishments in stochastic optimization, financial mathematics, and behavioral finance.

Chair's Column

Welcome again to the latest issue of *Views and News*. It has been an incredibly successful year for our SIAG and I'm happy to report that our charter, which was up for renewal, has been approved by the SIAM Board of Trustees. In addition, over the last 12 months our membership has grown from 961 to 1,091. Other activities that our SIAG has engaged in are organizing a SIAG/OPT track of sessions at the 2016 SIAM Annual Meeting, July 11–15, Boston, Massachusetts and a *SIAM News* article on Energy Optimization (Rain Is Free, or Isn't It?). I highly recommend reading it.

We're hoping to keep the momentum going for the rest of 2016 and 2017. As you know, we are well on our way towards our triennial conference, which will be held in Vancouver, British Columbia on May 22–25, 2017. The conference is shaping up nicely and I'll report on further developments in the next Views and News. In the meantime, I'm excited to announce that the conference will feature two minitutorials so you may want to check out those as well. The first is on Stochastic Optimization for Machine Learning, organized by Francis Bach and Mark Schmidt. The second will be on Optimal Power Flow and is being organized by Alper Atamtürk and Daniel Bienstock.

I also wanted to send out a friendly reminder that SIAG elections are just around the corner. We have a great slate of candidates and I urge all members to vote. The new officers will be announced in January. Please check your inbox for more information and how to go about voting.

Amidst all the great news, I'm saddened to report that we have recently lost some exceptional optimization researchers this year: Jonathan Borwein, Roger Fletcher, and Christodoulos Floudas. In addition to being outstanding researchers all three were also wonderful people always willing to help out others. We will miss them greatly.

As always, if you have any suggestions for activities that this SIAG can engage in to benefit its members, please feel free to contact any of the officers. I wish you all a wonderful rest of the year.

Juan Meza, SIAG/OPT Chair
School of Natural Sciences, University of California, Merced, Merced, CA 95343, USA, jcmeza@ucmerced.edu, http://bit.ly/1G8HUxO

Comments from the Editors

We thank the contributors to the 34th issue of SIAM Activity Group on Optimization's newsletter and congratulate the newly-elected SIAM fellows and prize winners!

Given the overwhelming interest in all things machine learning, Katya's article is a particularly timely take on the (forgive the pun) deep connections between optimization and ML.

We are served well by looking to the past, and this issue's personal remembrances by Adrian, Sven, and Ruth provide great insights into three luminaries of our field. We extend our condolences to the family, friends, and colleagues of Jon Borwein, Roger Fletcher, and Chris Floudas.

With election season in full swing in the U.S., it is interesting to remember that news coverage was not always what is currently broadcast: On Thanksgiving day in 1997, the cable news network MSNBC (then only a year old) broadcast a "pi(e)" feature with Jon Borwein, Peter Borwein, and Simon Plouffe from the campus of Simon Fraser University! Other remembrances can be found at the memorial site for Jon Borwein at http://jonborwein.org.

It is also worthwhile to share related activities happening in other SIAM activity groups. The SIAG/CSE has recently produced an extensive report on research and education in computational science and engineering, a discipline that was near and dear to Chris Floudas.

Many of you have written to opt for an electronic copy of *Views and News*; for the others among you, please do not hesitate to contact us to opt out of receiving physical copies. We cannot help to notice that Roger Fletcher was a master of minimizing paper: the two landmark papers on the DFP [59] and Fletcher-Reeves [60] methods were a combined 12 (11 if one allows for packing!) pages in length. His succinctness is something for us to aspire to, even in the digital age.

We welcome your feedback, (e-)mailed directly to us or to siagoptnews@lists.mcs.anl.gov. Suggestions for new issues, comments, and papers are always welcome! Best wishes for the new year, see you in 2017,

Stefan Wild, Editor
Mathematics and Computer Science Division, Argonne National Laboratory, USA, wild@anl.gov, http://www.mcs.anl.gov/~wild
Jennifer Erway, Editor
Department of Mathematics, Wake Forest University, USA, erwayjb@wfu.edu, http://www.wfu.edu/~erwayjb