Computing Gradients in Large-Scale Optimization Using Automatic Differentiation

Christian Bischof Ali Bouaricha
Peyvand Khademi Jorge Moré

CRPC-TR95582
June 1995

Center for Research on Parallel Computation
Rice University
6100 South Main Street
CRPC - MS 41
Houston, TX 77005

Also available as *MCS-P488-0195* from Argonne National Laboratory, Mathematics and Computer Science Division.

ARGONNE NATIONAL LABORATORY
9700 South Cass Avenue
Argonne, Illinois 60439

COMPUTING GRADIENTS IN LARGE-SCALE OPTIMIZATION USING AUTOMATIC DIFFERENTIATION

Christian H. Bischof, Ali Bouaricha, Peyvand M. Khademi, Jorge J. Moré

Mathematics and Computer Science Division

Preprint MCS-P488-0195

January 1995

(Revised version)

June 1995

Work supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Computational and Technology Research, U.S. Department of Energy, under Contract W-31-109-Eng-38, by the National Aerospace Agency under Purchase Order L25935D, and by the National Science Foundation, through the Center for Research on Parallel Computation, under Cooperative Agreement No. CCR-9120008.

COMPUTING GRADIENTS IN LARGE-SCALE OPTIMIZATION USING AUTOMATIC DIFFERENTIATION*

Christian H. Bischof, Ali Bouaricha, Peyvand M. Khademi, Jorge J. Moré

Abstract

The accurate and efficient computation of gradients for partially separable functions is central to the solution of large-scale optimization problems, since these functions are ubiquitous in large-scale problems. We describe two approaches for computing gradients of partially separable functions via automatic differentiation. In our experiments we employ the ADIFOR (Automatic Differentiation of Fortran) tool and the SparsLinC (Sparse Linear Combination) library. We use applications from the MINPACK-2 test problem collection to compare the numerical reliability and computational efficiency of these approaches with hand-coded derivatives and approximations based on differences of function values. Our conclusion is that automatic differentiation is the method of choice, providing code for the efficient computation of the gradient without the need for tedious hand-coding.

1 Introduction

The solution of nonlinear optimization problems often requires the computation of the gradient ∇f_0 of a mapping $f_0 : \mathbb{R}^n \to \mathbb{R}$. If the number of variables n is moderate, we can approximate the components of the gradient by differences of function values, for example,

$$[\nabla f_0(x)]_i \approx \frac{f(x + h_i e_i) - f(x)}{h_i}, \qquad 1 \le i \le n, \tag{1.1}$$

where h_i is the difference parameter, and e_i is the i-th unit vector. However, for large-scale problems (even for moderately sized problems with $n = 100$ variables) use of this approximation is prohibitive because it requires n function evaluations for each gradient. Another reason to avoid the use of (1.1) is that truncation errors in this calculation can mislead an optimization algorithm and cause premature termination far away from a solution. Thus, algorithms for the solution of optimization problems avoid approximations of the gradient by differences, and insist on an accurate and efficient evaluation of the gradient.

In this paper we explore the use of automatic differentiation tools for the computation of ∇f_0 when $f_0 : \mathbb{R}^n \mapsto \mathbb{R}$ is partially separable, that is, f_0 can be represented in the form

$$f_0(x) = \sum_{i=1}^{m} f_i(x), \tag{1.2}$$

*Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439-4843. This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Computational and Technology Research, U.S. Department of Energy, under Contract W-31-109-Eng-38, by the National Aerospace Agency under Purchase Order L25935D, and by the National Science Foundation, through the Center for Research on Parallel Computation, under Cooperative Agreement No. CCR-9120008.

where f_i depends on $p_i \ll n$ variables. This class of functions, introduced by Griewank and Toint [21, 22], plays a fundamental role in the solution of large-scale optimization problems since, as shown by Griewank and Toint, a function f_0 is partially separable if the Hessian matrix $\nabla^2 f_0(x)$ is sparse.

Algorithms and software that take advantage of the partially separable structure have been developed for various problems. See, for example, [27, 14, 23, 31, 32, 33, 34]. In these algorithms the partially separable structure is used mainly to approximate the (dense) Hessian matrices $\nabla^2 f_i(x)$ by quasi-Newton methods. Partial separability is also used to compute the gradient of f_0 as the sum of the gradients of the element functions f_i, but this is just another method for hand-coding the gradient. In a related paper [11] we discuss the impact of partial separability on optimization software.

The key observation needed to compute the gradient of a partially separable function is that if $f_0 : \mathbb{R}^n \mapsto \mathbb{R}$ is defined by (1.2), and if the vector-valued function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ defined by

$$f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}, \tag{1.3}$$

then the gradient of f_0 is given by

$$\nabla f_0(x) = f'(x)^T e, \tag{1.4}$$

where $f'(x)$ is the Jacobian matrix of f at x, and $e \in \mathbb{R}^m$ is the vector of all ones. At first sight this approach does not look promising because it requires the computation of the Jacobian matrix $f'(x)$. However, for partially separable functions, f_i depends on $p_i \ll n$ variables, and thus $f'(x)$ is a sparse matrix. We use the sparsity of $f'(x)$ to show that automatic differentiation tools can compute the gradient ∇f_0 so that

$$T\{\nabla f_0(x)\} \leq \Omega_T \, T\{f_0(x)\}, \tag{1.5}$$

$$M\{\nabla f_0(x)\} \leq \Omega_M \, M\{f_0(x)\}, \tag{1.6}$$

where $T\{\cdot\}$ and $M\{\cdot\}$ denote computing time and memory, respectively, and Ω_T and Ω_M are constant. We also show that for partially separable functions that arise in applications, the constants Ω_T and Ω_M are small and independent of n, in contrast to the use of (1.1).

The approach for computing the gradient of f_0 using (1.3) and (1.4) was proposed by Andreas Griewank and can be viewed as a special case of the results discussed by Griewank [18, Section 2]. Preliminary tests of this approach were done by Bischof and El-Khadiri [10]. The results in this paper show that this approach is not only feasible, but highly efficient.

A brief review of automatic differentiation, the ADIFOR (Automatic Differentiation of Fortran) tool [4, 6], and the SparsLinC (Sparse Linear Combination) library [5, 6] is

provided in the next section. Automatic differentiation techniques rely on the fact that every function, no matter how complicated, is executed on a computer as a potentially long sequence of elementary operations such as additions, multiplications, and elementary functions (e.g., the trigonometric and exponential functions). By applying the chain rule to the composition of those elementary operations, derivative information can be computed exactly and in a completely mechanical fashion [19, 28].

In Section 3 we propose two approaches for the computation of the Jacobian matrix $f'(x)$. The first approach uses the sparsity pattern of $f'(x)$, graph-coloring techniques, and the ADIFOR tool to obtain a compressed Jacobian matrix that contains all the information needed to determine the entire Jacobian matrix. The second approach uses ADIFOR with the SparsLinC library to produce a sparse representation of the Jacobian matrix without a priori knowledge of the sparsity pattern. In fact, the sparsity pattern is a byproduct of the ADIFOR/SparsLinC approach.

Section 4 discusses the formulation of large-scale problems in terms of partially separable functions, and outlines the problems from the MINPACK-2 [1] collection of large-scale problems that we use to validate our approach. Experimental results with problems from the MINPACK-2 collection on Sun SPARC 10, IBM RS 6000 (model 370), and Cray C90 platforms are presented in Section 5.

Our results show that the compressed Jacobian approach with the ADIFOR automatic differentiation tool generally outperforms difference approximations (to the compressed Jacobian matrix) in terms of computing time. The ADIFOR/SparsLinC approach obviates the need for the computation of the sparsity pattern and the compressed Jacobian matrix, but produces slower gradient code in our test problems. This tradeoff between convenience and cost is not always an option. Using ADIFOR/SparsLinC is the only feasible approach for applications where it is desirable to relieve the user of the error-prone task of providing the sparsity pattern or where the assumption that the sparsity pattern of $f'(x)$ is independent of x does not hold. For both approaches, (1.5) and (1.6) hold with constants Ω_T and Ω_M that are small and independent of n.

In terms of accuracy, both approaches provide the gradient to full accuracy, while approximations based on differences always suffer from truncation errors and provide, at best, half the accuracy in the function evaluation. We emphasize that the accuracy of the gradient in an optimization algorithm is of paramount importance because the gradient is used to determine the search directions. An inaccurate gradient can easily lead to false convergence.

2 The ADIFOR Tool and the SparsLinC Library

Automatic differentiation [19, 28] (AD) is a chain-rule-based technique for evaluating the derivatives of functions defined by computer programs. AD produces code that, in the absence of floating-point exceptions, computes the values of the analytical derivatives ac-

curate to machine precision. AD avoids the truncation and cancellation errors inherent in approximations of derivatives by differences of function values. Moreover, unlike symbolic approaches such as Maple, Macsyma, or Reduce, it is applicable to codes of arbitrary length containing branches, loops, and subroutine calls.

The *forward* and *reverse* modes of automatic differentiation for computing the Jacobian matrix of a mapping $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ are distinguished by how the chain rule is used to propagate derivatives through the computation. The forward mode accumulates the derivatives of intermediate variables with respect to the *independent* variables x, whereas the reverse mode propagates the derivatives of the *dependent* variables $y = f(x)$ with respect to intermediate variables.

Given a *seed* matrix S with n rows and p columns, the forward mode generates code for the computation of the directional derivative

$$f'(x)S. \tag{2.1}$$

The complexity of the forward mode is rather predictable. If $L\{f\}$ and $M\{f\}$ are, respectively, the number of floating-point operations and the amount of memory required by the computation of $f(x)$, then an AD-generated code employing the forward mode requires

$$L\{f'(x)S\} \leq (2+3p)\, L\{f\}, \qquad M\{f'(x)S\} \leq (1+p)\, M\{f\}.$$

floating-point operations and memory, respectively, to compute $f'(x)S$ (see Griewank [18]). With the reverse mode, on the other hand, we can compute $f'(x)^T Q$ where, Q is a seed matrix with m rows and q columns. The reverse mode requires the ability to reverse the partial order of program execution and to remember (or recompute) any intermediate result that nonlinearly affects the final result. As a result, the complexity of the reverse mode is harder to predict. If no intermediate values are recomputed, a straightforward implementation of the reverse mode requires $\mathcal{O}(L\{f\})$ floating-point operations and up to $\mathcal{O}(L\{f\} + M\{f\})$ memory, depending on the nonlinearity of the code.

The reverse mode is attractive when m is small. In particular, if $m = 1$, then $f'(x)$ is a gradient, and the reverse mode needs only $\mathcal{O}(L\{f\})$ operations to compute $f'(x)$. The storage requirement of the reverse mode, however, can be a difficulty because of the possible dependence on $L\{f\} + M\{f\}$. Griewank [17] suggested a snapshot approach to circumvent this difficulty.

There have been various implementations of automatic differentiation; an extensive survey can be found in [25]. In particular, we mention GRESS [24], and PADRE-2 [26] for Fortran programs and ADOL-C [20] for C programs. GRESS, PADRE-2, and ADOL-C implement both the forward and reverse modes. In order to save control flow information and intermediate values, these tools generate a trace of the computation by recording the particulars of every operation performed in the code. The interpretation overhead associated with

using this trace for the purposes of automatic differentiation, as well as its potentially very large size, can be a serious computational bottleneck [30]. Recently, a source transformation approach to automatic differentiation has been explored in the ADIFOR [4, 6], ADIC [9], AMC [16], and Odyssee [29] tools. ADIFOR transforms Fortran 77 code, ADIC transforms ANSI-C code, and AMC and Odyssee transform a subset of Fortran 77. ADIFOR and ADIC mainly use the forward mode, with the reverse mode at the statement level, while AMC and Odyssee use the reverse mode.

In our work, we employed the ADIFOR tool, which has been developed jointly by Argonne National Laboratory and Rice University.* Given a Fortran subroutine (or collection of subroutines) describing a function, and an indication of which variables in parameter lists or common blocks correspond to independent and dependent variables with respect to differentiation, ADIFOR produces Fortran 77 code that allows the computation of the derivatives of the dependent variables with respect to the independent variables.

The workhorse of any mainly forward-mode first-order automatic differentiation approach, such as employed in ADIFOR or ADIC, for computing the m directional derivatives in (2.1) is the vector linear combination

$$\sum_{i=1}^{k} \alpha_i v_i, \tag{2.2}$$

where α_i is a scalar, v_i is a vector of length p, and k is usually less than 10. By default, this operation is implemented as a `DO` loop; and as long as p is of moderate size and the vectors are dense, this is an efficient way of expressing a vector linear combination.

The SparsLinC library [5, 6] addresses the situation where the seed matrix S is sparse and most of the vectors involved in the computation of $f'(x)S$ are sparse. This situation arises, for example, in the computation of large sparse Jacobian matrices, since the sparsity of the final Jacobian matrix implies that, with great probability, all intermediate derivative computations involve sparse vectors as well. SparsLinC implements routines for executing the vector linear combination (2.2) using sparse data structures [6]. It is fully integrated into ADIFOR and ADIC and provides a mechanism for transparently exploiting sparsity in derivative computations. SparsLinC does not require knowledge of the sparsity structure of the Jacobian matrix; indeed, the sparsity structure of the Jacobian matrix is a byproduct of the derivative computation. The SparsLinC routines adapt to the particular situation at hand, providing efficient support for a wide variety of sparsity scenarios.

*See the World Wide Web site `http://www.mcs.anl.gov/Projects/autodiff/index.html` for additional information on ADIFOR and ADIC.

3 Computing Gradients of Partially Separable Functions

We compute the gradient of a partially separable function as outlined in Section 1: Given the element functions $f_1, \ldots, f_m$ that define the partially separable function (1.2), we compute the Jacobian matrix $f'(x)$ of the vector-valued function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ defined by (1.3). The gradient $\nabla f_0(x)$ of the partially separable function is then obtained via (1.4); that is, we add the rows of $f'(x)$. In this section we propose two techniques for computing the Jacobian matrix.

If the sparsity pattern of $f'(x)$ is known, then graph-coloring techniques can be used to determine a seed matrix S so that the *compressed Jacobian matrix* $f'(x)S$ contains all the information needed to determine the entire Jacobian matrix $f'(x)$. The compressed Jacobian matrix approach has long been used in connection with the determination of sparse Jacobian matrices by differences of function values; see, for example, [13, 15]. The compressed Jacobian matrix approach requires the determination of a partitioning of the columns of $f'(x)$ into *structurally orthogonal* columns, that is, columns that do not have a nonzero in the same row position. Because of the structural orthogonality property we can uniquely extract all entries of the original Jacobian matrix from the compressed Jacobian.

The partitioning problem can be considered as a graph-coloring problem [13]. Given a graph representation of the sparsity structure of $f'(x)$, these algorithms produce a partitioning of the columns of $f'(x)$ into p structurally orthogonal groups by graph-coloring algorithms for the column-intersection graph associated with $f'(x)$. For many sparsity patterns, p is small and independent of n. For example, if a matrix is banded with bandwidth β or if it can be permuted to a matrix with bandwidth β, it can be shown [13] that $p \leq \beta$. In our experiments we employ the graph-coloring software described in [12] to determine an appropriate partition.

In an optimization algorithm we invariably need to compute a sequence $\{\nabla f_0(x_k)\}$ of gradients for some sequence $\{x_k\}$ of iterates. This step requires the computation of a sequence of Jacobian matrices $\{f'(x_k)\}$. In most cases we need to do the graph-coloring only once, since we can specify the *closure* of the sparsity pattern, that is, a sparsity pattern that, for every iterate x_k, contains the sparsity pattern of $\{f'(x_k)\}$. If we are not able to specify the closure of the sparsity pattern, the compressed Jacobian approach requires a call to the graph-coloring software at each iteration.

By exploiting the capability to compute directional derivatives (2.1), compressed Jacobian matrices can easily be computed via automatic differentiation (for additional details, see [3]): Given the seed matrix S, ADIFOR computes the compressed Jacobian matrix $f'(x)S$. In contrast to the approximation techniques based on the compressed Jacobian matrix approach [13, 15], all columns of the compressed Jacobian matrix are computed at once.

In many situations it is desirable to have a tool for the determination of $f'(x)$ that does

not require knowledge of the sparsity pattern of $f'(x)$. This situation arises, for example, while developing interfaces to the solution of large-scale optimization problems [11], where it is desirable to relieve the user of the error-prone task of providing the sparsity pattern. In these situations, a sparse implementation of automatic differentiation, such as provided by the ADIFOR/SparsLinC approach, is the only feasible approach.

We use the term *sparse* ADIFOR for the approach based on the ADIFOR tool employing the SparsLinC library for the computation of vector linear combinations of derivative objects. This approach is extremely simple. We run ADIFOR with instructions to generate calls to SparsLinC. Then, at runtime, we set the seed matrix S to the identity matrix using the SparsLinC interface routines. No knowledge of the sparsity structure is required. On the other hand, this approach is likely to be slower than the compressed Jacobian approach because of the need to maintain dynamic data structures for the representation of the sparse vectors. We also note that, unlike the compressed Jacobian matrix approach, this approach is applicable to Jacobian matrices that have a few dense rows; SparsLinC will allocate a few long vectors for the dense rows and will maintain all others as short vectors.

4 Test Problems

A wide variety of large-scale optimization problems in applications can be formulated as variational problems where we need to minimize a functional of the form

$$\int_{\mathcal{D}} \Phi(x, v, \nabla v)\, dx, \tag{4.1}$$

where $\mathcal{D}$ is some domain in $\mathbb{R}^2$, and Φ is defined by the application. In all cases (4.1) is well defined if $v : \mathcal{D} \mapsto \mathbb{R}^p$ belongs to $H^1(\mathcal{D})$, the Hilbert space of functions such that v and $\|\nabla v\|$ belong to $L^2(\mathcal{D})$.

Finite element approximations to these problems are obtained by minimizing (4.1) over the space of piecewise linear function v with values $v_{i,j}$ at $z_{i,j}$, $0 \le i \le n_y+1$, $0 \le j \le n_x+1$, where $z_{i,j} \in \mathbb{R}^2$ are the vertices of a triangulation of $\mathcal{D}$ with grid spacings h_x and h_y. The vertices $z_{i,j}$ are chosen to be a regular lattice so that there are n_x and n_y interior grid points in the coordinate directions, respectively. Lower triangular elements T_L are defined by vertices $z_{i,j}, z_{i+1,j}, z_{i,j+1}$, while upper triangular elements T_U are defined by vertices $z_{i,j}, z_{i-1,j}, z_{i,j-1}$. A typical triangulation is shown in Figure 4.1.

The finite element approximation to (4.1) is defined by the values $v_{i,j}$ of a piecewise linear functions at $z_{i,j}$. The values $v_{i,j}$ are obtained by solving the minimization problem

$$\min \left\{ \sum_{(i,j)} \left(f_{i,j}^L(v) + f_{i,j}^U(v) \right) : v \in \mathbb{R}^n \right\}, \tag{4.2}$$

where $f_{i,j}^L$ and $f_{i,j}^U$ are the finite element approximation to the integrals in the elements T_L

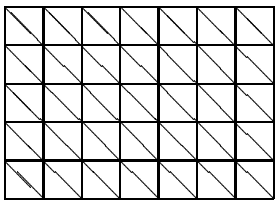

Figure 4.1: Triangulation of domain $\mathcal{D}$

and T_U, respectively. This problem can be expressed in partially separable form by setting

$$f(v) = \begin{pmatrix} f^L_{1,1}(v) \\ f^L_{1,2}(v) \\ \vdots \\ f^U_{1,1}(v) \\ f^U_{1,2}(v) \\ \vdots \end{pmatrix}. \tag{4.3}$$

The Jacobian matrix of this mapping is sparse, since the element functions $f^L_{i,j}(v)$ and $f^U_{i,j}(v)$ depend only on $v_{i,j}, v_{i+1,j}, v_{i,j+1}$ and $v_{i,j}, v_{i-1,j}, v_{i,j-1}$, respectively, and thus the techniques presented in Section 2 are directly applicable to the computation of the Jacobian matrix of this mapping.

There are other ways to express problem (4.2) in partially separable form. For example, by accumulating the contributions of the lower triangular elements T_L and T_U, we obtain the mapping

$$f(v) = \begin{pmatrix} f^L_{1,1}(v) + f^U_{1,1}(v) \\ f^L_{1,2}(v) + f^U_{1,2}(v) \\ \vdots \end{pmatrix}. \tag{4.4}$$

A difference between formulations (4.3) and (4.4) is that the number of element functions $m \approx 2n$ for (4.3), while $m \approx n$ for (4.4). This implies, in particular, that the number of groups p determined by the graph-coloring software is likely to be different, and thus the computing times for the compressed Jacobian matrix may depend on p. In our preliminary experience, however, the computing time of different formulations did not differ significantly.

We selected six problems from the MINPACK-2 test problem collection to compare the different approaches for computing the gradient of a partially separable function. The selected problems are representative of large-scale optimization problems arising from applications in superconductivity, optimal design, combustion, and lubrication. We give only a brief description of two of these problems to illustrate the partially separable structure of these problems. For further information refer to [1].

The Ginzburg-Landau (GL2) problem is of the form (4.1), where $v : \mathbb{R}^2 \mapsto \mathbb{R}^4$. The first two components of v represent a complex-valued function $\psi : \mathcal{D} \mapsto \mathbb{C}$ (the order parameter), and the other two components a vector-valued function $A : \mathcal{D} \mapsto \mathbb{R}^2$ (the vector potential). This problem has the form

$$\min\{f_1(\psi) + f_2(\psi, A) : \psi, A \in H_0^1(\mathcal{D})\},$$

where $\mathcal{D}$ is a two-dimensional region,

$$f_1(\psi) = \int_{\mathcal{D}} \left\{ -|\psi(x)|^2 + \tfrac{1}{2}|\psi(x)|^4 \right\} \, dx,$$

$$f_2(\psi, A) = \int_{\mathcal{D}} \left\{ \left\| [\nabla - iA(x)]\, \psi(x) \right\|^2 + \kappa^2 \left\| (\nabla \times A)(x) \right\|^2 \right\} \, dx,$$

and κ is the Ginzburg-Landau constant.

The minimal surface area (MSA) problem is of the form

$$\min\{f(v) : v \in K\},$$

where $f : K \mapsto \mathbb{R}$ is the functional

$$f(v) = \int_{\mathcal{D}} \left(1 + \|\nabla v(x)\|^2 \right)^{1/2} \, dx,$$

and the set K is defined by

$$K = \left\{ v \in H^1(\mathcal{D}) : v(x) = v_D(x) \text{ for } x \in \partial\mathcal{D} \right\}$$

for the boundary data function $v_D : \partial\mathcal{D} \mapsto \mathbb{R}$ that specifies the Enneper minimal surface.

These two problems are partially separable, but each code is structured distinctly, resulting in a distinctly structured compressed Jacobian in each case (the other four MINPACK-2 problems, SSC, EPT, ODC, and PJB, are all structurally identical to the MSA problem). For the GL2 problem (where $p = 8$), the compressed Jacobian turns out to be 50% dense, whereas for the MSA problem (where $p = 3$), the compressed Jacobian is almost completely dense. As we shall see, respectively in Sections 5.2 and 5.3, this variance in densities impacts the memory requirements and computing time performance of the ADIFOR/SparsLinC approach relative to that of the ADIFOR approach.

5 Experimental Results

We compare four methods for the computation of the gradient of a partially separable function: hand-coded derivative (HC), approximation of the compressed Jacobian matrix with function differences (FD), computation of the compressed Jacobian matrix with ADIFOR (AD), and computation of the full Jacobian matrix with ADIFOR/SparsLinC (Sparse AD).

Our aim is to compare these methods with the cost of computing the function (F) and to show that in all cases (1.5) and (1.6) hold with constants Ω_T and Ω_M that are small and independent of n.

Experiments were performed on Sun SPARC 10, an IBM RS 6000 (model 370), and a Cray C90. The Fortran compiler was used with all optimization options turned on.[†] All computations were done with 64-bit arithmetic.

The MINPACK-2 problems were used as a test set because the availability of hand-coded gradients provides a metric in terms of accuracy, computing time, and memory requirements. The emphasis of our work is to show the effectiveness of automatic differentiation tools for computing gradients, given that for many problems hand-coding of derivatives is non trivial and prohibitive in cost.

5.1 Numerical Accuracy

In terms of numerical accuracy, the approaches based on automatic differentiation were accurate to near machine precision, while the approach based on function differences were accurate up to at most half of the number of possible significant digits. We do not elaborate further on this point because this contrast in accuracies between automatic differentiation and function differences shows consistency with previously published work [3] on the computation of sparse Jacobian matrices with automatic differentiation.

5.2 Memory Requirements

Tables 5.1 and 5.2 present, respectively for the GL2 and MSA problems, the total memory required for the computation of the function as well as the various gradient methods, for the case of $n = 160,000$ variables. The remaining four problems have identical memory requirements to each other; these are shown in Table 5.3.

We measured memory with the Unix command `size` *executable-file*, which reports the total amount of statically allocated memory (memory requirements that can be assessed at compile time) needed to load and run the executable. In the case of SparsLinC, where memory is also allocated dynamically, we call a SparsLinC routine that reports the total amount of dynamically allocated memory, and we add this to the statically allocated memory.

The AD and FD approaches have similar memory requirements for the gradient computation. In both cases, memory requirements for the compressed Jacobian matrix are proportional to the product mp, where m is the number of component functions of f, and p is the number of groups determined by the graph-coloring algorithm. Sparsity pattern and graph-coloring computations, present in both approaches, require memory proportional to

[†]On the Sun, we employed F77 version 1.4 with the -O option; on the IBM, XLF version 3.1.2.3 with the -O option; and on the Cray, CFT77 version 6.0.3.20 with the -O inline3 -Oscalar3 -O task0 -O vector3 -Wf"-dp" options.

Table 5.1: Memory Requirements for GL2 (in Mbytes; $n = 160,000$)

Platform	F	FD	FD/F	AD	AD/F	Sparse AD	Sparse AD/F
SPARC / IBM	2.59	31.39	12.1	48.13	18.6	38.65	15.0
Cray C90	3.07	42.76	13.9	59.50	19.4	59.74	19.5

Table 5.2: Memory Requirements for MSA (in Mbytes; $n = 160,000$)

Platform	F	FD	FD/F	AD	AD/F	Sparse AD	Sparse AD/F
SPARC / IBM	2.57	34.68	13.5	33.38	13.0	39.64	15.4
Cray C90	2.99	49.19	16.5	47.90	16.0	60.37	20.2

Table 5.3: Memory Requirements for SSC, EPT, ODC, or PJB (in Mbytes; $n = 160,000$)

Platform	F	FD	FD/F	AD	AD/F	Sparse AD	Sparse AD/F
SPARC / IBM	1.29	33.38	25.8	32.09	24.8	38.55	29.9
Cray C90	1.72	47.91	27.9	46.61	27.1	59.39	34.9

$\mathrm{nnz}(f'(x))$, the total number of nonzeros in the Jacobian matrix. Each approach also has some distinct memory requirements which account for the differences between the two in Tables 5.1–5.3.

For the Sparse AD approach, much of the memory is allocated dynamically and based on the need to represent nonzero derivative information. Certainly, the memory needed for representing the sparse Jacobian matrix has a lower bound of $\mathrm{nnz}(f'(x))$. Beyond this, SparsLinC requires additional memory for internal representations as explained in [8].

The first column in Tables 5.1–5.3 shows the memory required for running the original function. Memory requirements for the hand-coded MINPACK-2 gradient codes are not shown separately, but are always between a factor of 1.5–2 times the memory requirements of the corresponding function. The next three double columns show the memory requirements of the FD, AD, and Sparse AD approaches in megabytes (Mbytes) and as the ratio of gradient to function memory requirements. The memory requirements on the SPARC 10 and IBM RS 6000 are identical, while the Cray C90 requires more memory because the Cray default length for integer variables is 64 bits, whereas it is 32 bits on the workstation platforms. This is particularly noticeable for the Sparse AD approach, which maintains integer arrays for sparse vector data structures.

The results in Tables 5.1–5.3 show that the strategy of computing the gradient of a partially separable function by reformulating the problem as the computation of the sparse Jacobian matrix of the mapping defined by (1.3) imposes modest memory requirements.

The memory requirements can also be measured in terms of the possible range of the constant Ω_M in (1.6). The table below shows that Ω_M is a small multiple of p. In these results we have rounded the coefficients of p to the nearest integer, since we are interested only in general trends.

	SPARC/IBM	Cray C90
FD	$p \le \Omega_M \le 9p$	$2p \le \Omega_M \le 9p$
AD	$2p \le \Omega_M \le 8p$	$2p \le \Omega_M \le 9p$
Sparse AD	$2p \le \Omega_M \le 10p$	$2p \le \Omega_M \le 12p$

All three approaches are comparable in terms of memory requirements. The worst performance is obtained for the problems in Table 5.3 because the function codes for these problems are relatively simple and require only the storage of the vector x. The results for the GL2 and MSA problems are more representative because these problems have work arrays in the function code. In general we expect the Sparse AD approach to require less memory than AD when the compressed Jacobian matrix is sparse. Indeed, the Sparse AD approach requires about 20% less memory on the workstation platforms for the GL2 problem, where the compressed Jacobian matrix is 50% sparse.

5.3 Computing Time

Figure 5.1 summarizes the GL2 and MSA results for the SPARC 10, IBM RS 6000 and Cray C90. Each figure shows the gradient-to-function computing time ratio for each of the four methods for computing the gradient. We have included data for problems with $n = 2,500$ variables to $n = 160,000$. The solid line indicates the Sparse AD approach, the dotted line the AD approach, the dashed line the FD approach, and the dash-dotted line is the hand-coded derivatives (HC).

The main conclusion that can be drawn from Figure 5.1 is that the gradient-to-function computing time ratio is independent of the problem size for these two problems. This is an important aspect of these results, since our main goal is to avoid the cost of n function evaluations for approximating the gradient by differences of function values. The gradient-to-function ratios for SparsLinC on the Cray C90 are not shown in Figure 5.1 because inclusion of these ratios would distort the plots. Table 5.4 show that these ratios, though larger, are also independent of n.

We are also interested in the ratio of computing times between the various approaches and their relation to the time required for the coloring preprocessing step. These ratios appear in Table 5.5 for all the problems under consideration, but only for $n = 160,000$. The plots in Figure 5.1 show that these ratios are essentially independent of the number n of variables, and thus the results in Table 5.5 are representative for any reasonable number of variables.

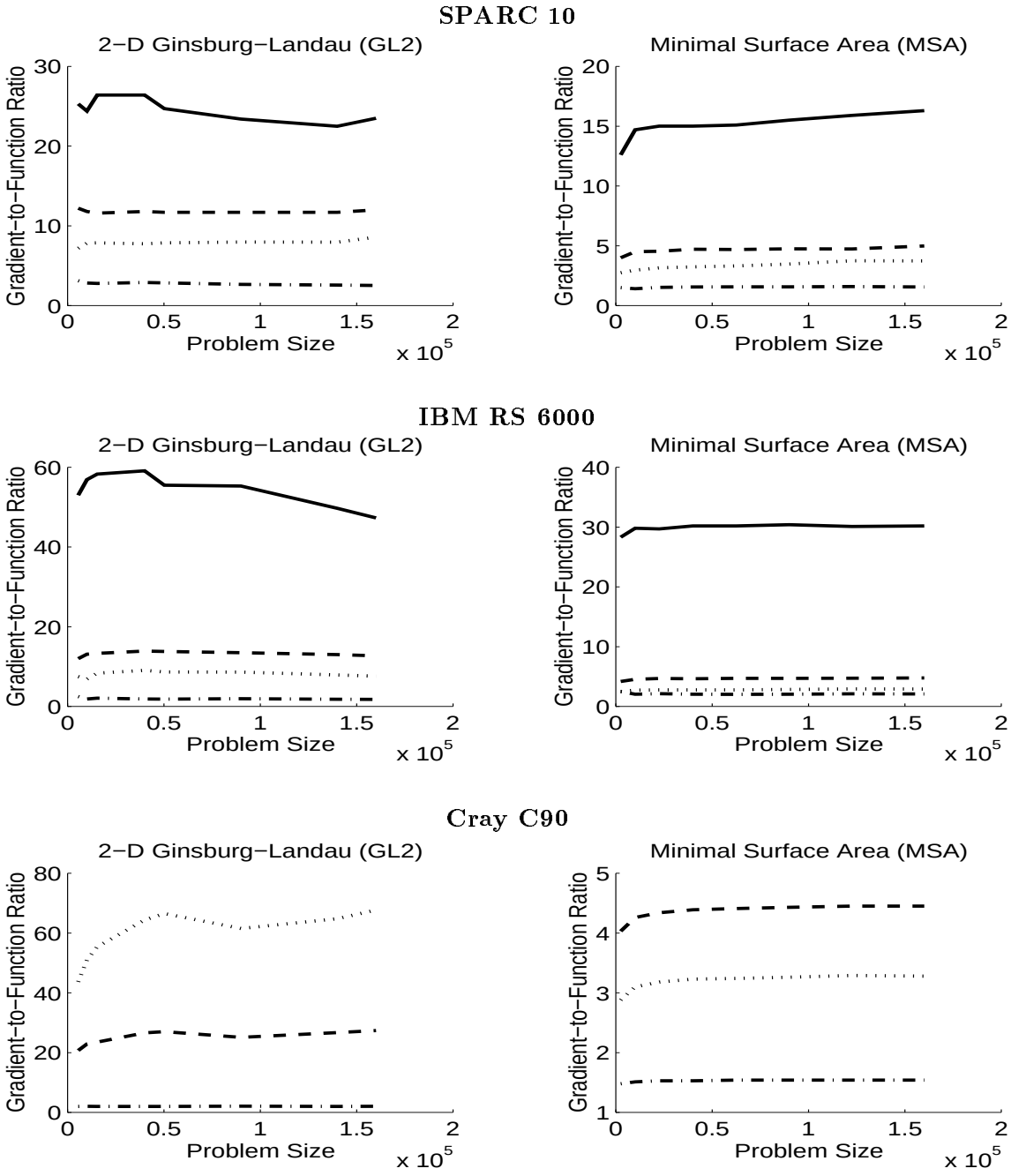

Figure 5.1: Ratios of computing times between the gradient and the function. FD (dashed), AD (dotted), Sparse AD (solid), HC (dash-dotted)

Table 5.4: Gradient-to-Function Runtime Ratios for Sparse AD on the Cray C90

n	10,000	40,000	90,000	160,000
GL2	1,390	1,790	1,710	1,790
MSA	70.7	72.1	72.2	72.6

Before we analyze the runtime results, we briefly summarize important features of the underlying architectures. The SPARC 10 essentially has a scalar processor and a flat memory hierarchy. Hence, vector operations execute only marginally faster, and memory locality (that is, the reuse of data and the accessing of adjacent memory locations) is not much of an issue. In contrast, the IBM RS 6000 architecture employs a superscalar chip and a cache-based memory architecture. Hence, this machine performs better if executing short vector operations, since these operations can fill the short pipes and take advantage of memory locality. On the other hand, indirect addressing, used extensively in SparsLinC and in the coloring algorithm, while fairly inconsequential on the SPARC, may lead to performance degradation, as memory locality suffers. The Cray C90 is a vector processor without a cache and achieves its full potential only when the code exhibits long vector operations. Without optimization of the source Fortran code, short vector loops and indirect addressing schemes exhibit much lower performance, since the hardware pipes cannot get filled and the speed of main memory is much slower than that of the CPU.

Based on the architectures used in our testing, we expect computing times to be stable and predictable on workstation platforms but expect that vectorization issues will cause a large variation in computing times on vector architectures. Our experimental results bear out these expectations.

As expected, the hand-coded derivative code is the fastest on the scalar architectures. For the results in Table 5.5 we have

$$T\{\nabla f_0(x) : \mathrm{HC}\} \leq 3\, T\{f_0\}, \tag{5.1}$$

where $T\{\nabla f_0(x) : \cdot\}$ is the time required to compute the gradient of the partially separable function by a particular method. The above ratio can be expected for well-coded gradient computations on scalar architectures but requires special techniques on vector and parallel architectures [2].

On vector architectures we can expect the ratio (5.1) to hold only if both the function and the gradient evaluation codes vectorize or if neither code vectorizes. An examination of Cray C90 results shows that only the MSA and ODC function evaluation codes fail to vectorize, and that the GL2 hand-coded gradient evaluation code is the only HC code that vectorizes. Our results support this remark because we obtain a high gradient-to-function

Table 5.5: Coloring-to-Function and Gradient-to-Function Runtime Ratios ($n = 160,000$)

Ginzburg-Landau (GL2) problem

Platform	Coloring	HC	FD	AD	Sparse AD
SPARC 10	18.36	2.52	12.00	8.58	23.50
IBM RS 6000	36.24	1.78	12.70	7.58	47.30
Cray C90	664.29	2.05	27.40	67.70	1790.00

Minimal Surface Area (MSA) problem

Platform	Coloring	HC	FD	AD	Sparse AD
SPARC 10	7.18	1.55	4.98	3.74	16.30
IBM RS 6000	11.62	2.09	4.77	2.90	30.20
Cray C90	12.57	1.54	4.45	3.28	72.60

Steady State Combustion (SSC) problem

Platform	Coloring	HC	FD	AD	Sparse AD
SPARC 10	4.43	1.28	4.63	3.08	18.00
IBM RS 6000	5.58	1.48	4.39	2.12	26.50
Cray C90	86.51	18.00	7.54	33.60	902.00

Optimal Design with Composites (ODC) problem

Platform	Coloring	HC	FD	AD	Sparse AD
SPARC 10	5.35	1.28	4.79	3.37	15.70
IBM RS 6000	7.68	1.43	4.55	2.56	26.30
Cray C90	10.25	2.09	4.41	4.95	77.60

Elastic-Plastic Torsion (EPT) problem

Platform	Coloring	HC	FD	AD	Sparse AD
SPARC 10	13.24	1.59	5.99	5.67	43.80
IBM RS 6000	23.88	2.50	5.71	4.46	87.70
Cray C90	331.98	25.5	17.50	63.30	2800.00

Pressure in a Journal Bearing (PJB) problem

Platform	Coloring	HC	FD	AD	Sparse AD
SPARC 10	12.64	1.92	5.82	5.06	25.20
IBM RS 6000	18.24	2.13	5.53	4.06	41.50
Cray C90	204.63	64.70	12.20	39.10	1260.00

runtime ratio only on problems where only the function evaluation code vectorizes (i.e., SSC, EPT, and PJB).

The results in Table 5.5 show that the AD approach outperforms the FD approach on scalar architectures. The performance of the various approaches on vector architectures is harder to predict as performance depends on the delicate interplay between the code and the compiler (for examples, see [7, 11]). Note that the results in Table 5.5 show that the performance of AD is comparable to that of FD on the Cray C90 for those problems (MSA and ODC) where the function evaluation code fails to vectorize.

Our numerical results also show that the AD approach outperforms the Sparse AD approach on all the architectures. From the results in Table 5.5 we can observe that

$$T\{\nabla f_0(x) : \text{ADIFOR}\} \leq \kappa\, T\{\nabla f_0(x) : \text{Sparse ADIFOR}\},$$

where κ satisfies

SPARC 10	IBM RS 6000	Cray C90
$3 \leq \kappa \leq 8$	$6 \leq \kappa \leq 20$	$15 \leq \kappa \leq 45$

.

In all our experiments with the exception of the GL2 problem, the compressed Jacobian is almost fully dense. It is not surprising that AD outperforms Sparse AD on these problems, given that the runtime efficiency of SparsLinC is expected to become apparent for problems that have much sparser compressed Jacobians. Note that Sparse AD performs much better on the GL2 problem, where the compressed Jacobian is 50% sparse, compared with the other problems.

We can compare the performance of the various approaches by computing the range for the constant Ω_T in (1.5) as a function of p. In these results we have also rounded the coefficients of p to the nearest integer.

	SPARC 10	IBM RS 6000	Cray C90
FD	$p \leq \Omega_T \leq 2p$	$p \leq \Omega_T \leq 2p$	$p \leq \Omega_T \leq 6p$
AD	$p \leq \Omega_T \leq 2p$	$p \leq \Omega_T \leq 2p$	$p \leq \Omega_T \leq 20p$
Sparse AD	$3p \leq \Omega_T \leq 15p$	$6p \leq \Omega_T \leq 30p$	$25p \leq \Omega_T \leq 930p$

The above table shows that in most cases Ω_T is a small multiple of p.

We note the wide variation in Ω_T for FD and AD on the vector architecture owing to the code-dependent effects of vectorization, as already discussed. We also note the large variation in Ω_T for the Sparse AD results on the SPARC 10. This results from the way SparsLinC exploits the particular sparsity characteristics of each problem (this issue is explored in [8]). Finally, we note that the performance of Sparse AD degrades on vector computers, as a result of pervasive use of indirect addressing and lack of vector instructions, though this performance could be improved through the use of hardware-supported gather/scatter instructions.

Table 5.5 also compares the cost of the graph coloring algorithm with the cost of computing the function. The high relative cost of computing the graph coloring is mainly a reflection of the low cost of computing the functions for these problems. We can justify this remark by noting that evaluation of the component functions for the GL2, EPT, and PJB problems only require the evaluation of a low-order polynomial and, that for these problems, the coloring-to-function runtime ratio is high. On the other hand, the problems with a low coloring-to-function runtime ratio are relatively expensive to evaluate; the SSC problem requires the evaluation of the exponential function, while the MSA and SSC problems require a square root.

Another reason for the high relative cost of computing the graph coloring is that the algorithm we employ (subroutine DSM from Coleman, Garbow, and Moré [12]) is intended to produce graph colorings with a small p by employing several heuristics. The runtime of subroutine DSM could be reduced by a factor of two or more without a substantial increase in p by only using one of the heuristics. Also note that the graph coloring algorithms share many of the characteristics of Sparse AD with respect to indirect addressing and memory locality, and thus the performance of the coloring algorithm deteriorates on the RS 6000 and C90 platforms.

6 Conclusions

We have shown that automatic differentiation outperforms difference approximations of derivatives and offers high numerical accuracy without the need for hand-coding. The approach based on the compressed Jacobian matrix with the ADIFOR tool produces code that is often not more than four times slower than a well-coded hand-derived gradient code on scalar architectures. This approach, however, requires the sparsity pattern of the partially separable function.

The approach based on the ADIFOR/SparsLinC tool set is the ultimate in convenience, as not even the sparsity pattern of the underlying Jacobian matrix is needed. In fact, the sparsity pattern is a byproduct of the ADIFOR/SparsLinC approach. On the other hand, this approach is considerably slower, particularly on vector architectures.

Acknowledgments

We thank Andreas Griewank for stimulating discussions on the subject and Alan Carle for his instrumental role in the ADIFOR project.

References

[1] B. M. Averick, R. G. Carter, J. J. Moré, and G.-L. Xue, *The* MINPACK-2 *test problem collection*, Preprint MCS-P153-0692, Mathematics and Computer Science

Division, Argonne National Laboratory, 1992.

[2] B. M. Averick and J. J. Moré, *Evaluation of large-scale optimization problems on vector and parallel architectures*, SIAM J. Optimization, 4 (1994), pp. 708–721.

[3] B. M. Averick, J. J. Moré, C. H. Bischof, A. Carle, and A. Griewank, *Computing large sparse Jacobian matrices using automatic differentiation*, SIAM J. Sci. Statist. Comput., 15 (1994), pp. 285–294.

[4] C. Bischof, A. Carle, G. Corliss, A. Griewank, and P. Hovland, *ADIFOR: Generating derivative codes from Fortran programs*, Scientific Programming, 1 (1992), pp. 11–29.

[5] C. Bischof, A. Carle, and P. Khademi, *Fortran 77 interface specification to the SparsLinC library*, Tech. Report ANL/MCS-TM-196, Mathematics and Computer Science Division, Argonne National Laboratory, 1994.

[6] C. Bischof, A. Carle, P. Khademi, and A. Mauer, *The ADIFOR 2.0 system for the automatic differentiation of Fortran 77 programs*, 1994. Preprint MCS-P481-1194, Mathematics and Computer Science Division, Argonne National Laboratory, and CRPC-TR94491, Center for Research on Parallel Computation, Rice University.

[7] C. Bischof, L. Green, K. Haigler, and T. Knauff, *Parallel calculation of sensitivity derivatives for aircraft design using automatic differentiation*, in Proceedings of the 5th AIAA/NASA/USAF/ISSMO Symposium on Multidisciplinary Analysis and Optimization, AIAA 94-4261, American Institute of Aeronautics and Astronautics, 1994, pp. 73–84.

[8] C. Bischof, P. Khademi, and A. Carle, *Fast computation of gradients and Jacobians by transparent exploitation of sparsity in automatic differentiation*, Preprint MCS-P519-0595, Mathematics and Computer Science Division, Argonne National Laboratory, 1995.

[9] C. Bischof and A. Mauer, *ADIC – A tool for the automatic differentiation of C programs*, Preprint MCS-P499-0295, Mathematics and Computer Science Division, Argonne National Laboratory, 1995.

[10] C. H. Bischof and M. El-Khadiri, *Extending compile-time reverse mode and exploiting partial separability in ADIFOR*, Tech. Report ANL/MCS-TM-163, Mathematics and Computer Science Division, Argonne National Laboratory, 1992.

[11] A. Bouaricha and J. Moré, *Impact of partial separability on large-scale optimization*, Preprint MCS-P487-0195, Mathematics and Computer Science Division, Argonne National Laboratory, 1995.

[12] T. F. Coleman, B. S. Garbow, and J. J. Moré, *Fortran subroutines for estimating sparse Jacobian matrices*, ACM Trans. Math. Software, 10 (1984), pp. 346–347.

[13] T. F. Coleman and J. J. Moré, *Estimation of sparse Jacobian matrices and graph coloring problems*, SIAM J. Numer. Anal., 20 (1983), pp. 187–209.

[14] A. R. Conn, N. I. M. Gould, and P. L. Toint, LANCELOT, Springer Series in Computational Mathematics, Springer-Verlag, 1992.

[15] A. R. Curtis, M. J. D. Powell, and J. K. Reid, *On the estimation of sparse Jacobian matrices*, J. Inst. Math. Appl., 13 (1974), pp. 117–119.

[16] R. Giering, *Adjoint model compiler, manual version 0.2, AMC version 2.04*, tech. report, Max-Planck Institut für Meteorologie, August 1992.

[17] A. Griewank, *Achieving logarithmic growth of temporal and spatial complexity in reverse automatic differentiation*, Optimization Methods and Software, 1 (1992), pp. 35–54.

[18] ——, *Some bounds on the complexity of gradients, Jacobians, and Hessians*, in Complexity in Nonlinear Optimization, P. Pardalos, ed., World Scientific Publishers, 1993, pp. 128–161.

[19] A. Griewank and G. F. Corliss, eds., *Automatic Differentiation of Algorithms: Theory, Implementation, and Application*, Society for Industrial and Applied Mathematics, 1991.

[20] A. Griewank, D. Juedes, and J. Srinivasan, *ADOL-C, a package for the automatic differentiation of algorithms written in C/C++*, Preprint MCS-P180-1190, Mathematics and Computer Science Division, Argonne National Laboratory, 1990.

[21] A. Griewank and P. L. Toint, *On the unconstrained optimization of partially separable functions*, in Nonlinear Optimization 1981, M. J. D. Powell, ed., Academic Press, 1982.

[22] ——, *Partitioned variable metric updates for large structured optimization problems*, Numer. Math., 39 (1982), pp. 119–137.

[23] ——, *Numerical experiments with partially separable optimization problems*, in Numerical Analysis: Proceedings Dundee 1983, D. F. Griffiths, ed., Lecture Notes in Mathematics 1066, Springer-Verlag, 1984.

[24] J. E. Horwedel, *GRESS: A preprocessor for sensitivity studies on Fortran programs*, in Automatic Differentiation of Algorithms: Theory, Implementation, and Application, A. Griewank and G. F. Corliss, eds., SIAM, Philadelphia, 1991, pp. 243–250.

[25] D. Juedes, *A taxonomy of automatic differentiation tools*, in Proceedings of the Workshop on Automatic Differentiation of Algorithms: Theory, Implementation, and Application, A. Griewank and G. Corliss, eds., Philadelphia, 1991, SIAM, pp. 315–330.

[26] K. Kubota, *PADRE2, a FORTRAN precompiler yielding error estimates and second derivatives*, in Automatic Differentiation of Algorithms: Theory, Implementation, and Application, A. Griewank and G. F. Corliss, eds., SIAM, Philadelphia, 1991, pp. 251–262.

[27] M. Lescrenier, *Partially separable optimization and parallel computing*, Ann. Oper. Res., 14 (1988), pp. 213–224.

[28] L. B. Rall, *Automatic Differentiation: Techniques and Applications*, vol. 120 of Lecture Notes in Computer Science, Springer Verlag, Berlin, 1981.

[29] N. Rostaing, S. Dalmas, and A. Galligo, *Automatic differentiation in Odyssee*, Tellus, 45a (1993), pp. 558–568.

[30] E. Soulie, *User's experience with Fortran compilers for least squares problems*, in Automatic Differentiation of Algorithms: Theory, Implementation, and Application, A. Griewank and G. F. Corliss, eds., SIAM, Philadelphia, 1991, pp. 297–306.

[31] P. L. Toint, *Numerical solution of large sets of algebraic nonlinear equations*, Math. Comp., 46 (1986), pp. 175–189.

[32] ——, *On large scale nonlinear least squares calculations*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 416–435.

[33] P. L. Toint and D. Tuyttens, *On large-scale nonlinear network optimization*, Math. Programming, 48 (1990), pp. 125–159.

[34] ——, *LSNNO: A Fortran subroutine for solving large-scale nonlinear network optimization problems*, ACM Trans. Math. Software, 18 (1992), pp. 308–328.