Using Programming Language Theory to Make Automatic Differentiation Sound and Efficient

Barak A. Pearlmutter[1] and Jeffrey Mark Siskind[2]

[1] Hamilton Institute, National University of Ireland Maynooth, Co. Kildare, Ireland
barak@cs.nuim.ie
[2] School of Electrical and Computer Engineering, Purdue University, 465 Northwestern Avenue, West Lafayette IN 47907-2035 USA qobi@purdue.edu

Summary. This paper discusses a new Automatic Differentiation (AD) system that correctly and automatically accepts nested and dynamic use of the AD operators, without any manual intervention. The system is based on a new formulation of AD as highly generalized first-class citizens in a λ-calculus, which is briefly described. Because the λ-calculus is the basis for modern programming-language implementation techniques, integration of AD into the λ-calculus allows AD to be integrated into an aggressive compiler. We exhibit a research compiler which does this integration. Using novel analysis techniques, it accepts source code involving free use of a first-class forward AD operator and generates object code which attains numerical performance comparable to, or better than, the most aggressive existing AD systems.

Key words: Nesting, lambda calculus, multiple transformation, forward mode, optimization

1 Introduction

Over sixty years ago, Church [1] described a model of computation which included higher-order functions as first-class entities. This λ*-calculus*, as originally formulated, did not allow AD operators to be defined, but Church did use the derivative operator as an example of a higher-order function with which readers would be familiar. Although the λ-calculus was originally intended as a model of computation, it has found concrete application in programming languages *via* two related routes. The first route came from the realization that extremely sophisticated computations could be expressed crisply and succinctly in the λ-calculus. This led to the development of programming languages (LISP, ALGOL, ML, SCHEME, HASKELL, etc.) that themselves embody the central aspect of the λ-calculus: the ability to freely create and apply functions including higher-order functions. The second route arose from the recognition that various program transformations and programming-language theoretic constructs were naturally expressed using the λ-calculus. This resulted in the use of the λ-calculus as the central mathematical scaffolding of programming-language theory (PLT): both as the formalism in which the semantics of programming-language constructs (conditionals, assignments, objects, exceptions, etc.) are mathematically defined, and as the intermediate format into which computer programs are converted for analysis and optimization.

A substantial subgroup of the PLT community is interested in advanced or functional programming languages, and has spent decades inventing techniques by which programming languages with higher-order functions can be made efficient. These techniques are part of the body of knowledge we refer to as PLT, and are the basis of the implementation of modern programming-language systems: JAVA, C♯, the GHC HASKELL compiler, GCC 4.x, etc. Some of these techniques are being gradually rediscovered by the AD community. For instance, a major feature in TAPENADE [2] is the utilization of a technique by which values to which a newly-created function refer are separated from the code body of the function; this method is used ubiquitously in PLT, where it is referred to as *lambda lifting* or *closure conversion* [4].

We point out that—*like it or not*—the AD transforms are higher-order functions: functions that both take and return other functions. As such, attempts to build implementations of AD which are efficient and correct encounter the same technical problems which have already been faced by the PLT community. In fact, the technical problems faced in AD are a superset of these, as the machinery of PLT, as it stands, is unable to fully express the reverse AD transformation. The present authors have embarked upon a sustained project to bring the tools and techniques of PLT—suitably augmented—to bear on AD. To this end, novel machinery has been crafted to incorporate first-class AD operators (functions that perform forward- and reverse-mode AD) into the λ-calculus. This solves a host of problems: (1) the AD transforms are specified formally and generally; (2) nesting of the AD operators, and inter-operation with other facilities like memory allocation, is assured; (3) it becomes straightforward to integrate these into aggressive compilers, so that AD can operate in concert with code optimization rather than beforehand; (4) sophisticated techniques can migrate various computations from run time to compile time; (5) a callee-derives API is supported, allowing AD to be used in a modular fashion; and (6) a path to a formal semantics of AD, and to formal proofs of correctness of systems that use and implement AD, is laid out.

Due to space limitations, the details of how the λ-calculus can be augmented with AD operators is beyond our scope. Instead, we will describe the basic intuitions that underly the approach, and exhibit some preliminary work on its practical benefits. This starts (Sect. 2) with a discussion of modularity and higher-order functions in a numerical context, where we show how higher-order functions can solve some modularity issues that occur in many current AD systems. We continue (Sect. 3) by considering the AD transforms as higher-order functions, and in this context we generalize their types. This leads us (Sect. 4) to note a relationship between the AD operators and the pushforward and pullback constructions of differential geometry, which motivates some details of the types we describe as well as some of the terminology we introduce. In Sect. 5 we discuss how constructs that appear to the programmer to involve run-time transforms can, by appropriate compiler techniques, be migrated to compile-time. Section 6 describes a system which embodies these principles. It starts with a minimalist language (the λ-calculus augmented with a numeric basis and the AD operators) but uses aggressive compilation techniques to produce object code that is competitive with the most sophisticated current FORTRAN-based AD systems. Armed with this practical benefit, we close (Sect. 7) with a discussion of other benefits which this new formalism for AD has now put in our reach.

2 Functional Programming and Modularity in AD

Let us consider a few higher-order functions which a numerical programmer might wish to use. Perhaps the most familiar is numerical integration,

```
double nint(double f(double), double x0, double x1);
```

which accepts a function $f : \mathbb{R} \to \mathbb{R}$ and range limits a and b and returns an approximation of $\int_a^b f(x)\,\mathrm{d}x$. In conventional mathematical notation we would say that this function has the type

$$\texttt{nint} : (\mathbb{R} \to \mathbb{R}) \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}.$$

There are a few points we can make about this situation.

First, note that the caller of `nint` might wish to pass an argument function which is not known, at least in its details, until run time. For example, in the straightforward code to evaluate

$$\sum_{i=1}^{n} \int_1^2 (\sin x)^{\cos(x/i)}\,\mathrm{d}x$$

the caller needs to make a function which maps $x \mapsto (\sin x)^{\cos(x/i)}$ for each desired value of i. Although it is possible to code around this necessity by giving `nint` a more complicated API and forcing the caller to package up this extra "environment" information, this is not only cumbersome and error prone but also tends to degrade performance. The notation we will adopt for the construction of a function, "closed" over the values of any relevant variables in scope at the point of creation, is a "λ expression," after which the λ-calculus is named. Here, it would be $(\lambda x\,.\,(\sin x)\hat{}(\cos(x/i)))$.

Second, note that it would be natural to define two-dimensional numerical integration in terms of nested application of `nint`. So for example,

```
double nint2(double f2(double x, double y),
             double x0, double x1,
             double y0, double y1)
{ return nint((λx . nint((λy . f(x,y)), y0, y1)),
              x0, x1); }
```

Similar nesting would occur, without the programmer being aware of it, if a seemingly-simple function defined in a library happened to use AD internally, and this library function were invoked within a function to which AD was applied.

Third, it turns out that programs written in functional-programming languages are rife with constructs of this sort (for instance, `map` which takes a function and a list and returns a new list whose elements are computed by applying the given function to corresponding elements of the original list); because of this, PLT techniques have been developed to allow compilers for functional languages to optimize across the involved procedure-call barriers. This sort of optimization has implications for numerical programming, as numerical code often calls procedures like `nint` in inner loops. In fact, benchmarks have shown the efficacy of these techniques on numerical code. For instance, code involving a double integral of this sort experienced an order of magnitude improvement over versions in hand-tuned FORTRAN or C, when written in SCHEME and compiled with such techniques (see `ftp://ftp.ecn.purdue.edu/qobi/integ.tgz` for details.)

Other numeric routines are also naturally viewed as higher-order functions. Numerical optimization routines, for instance, are naturally formulated as procedures which take the function to be optimized as one argument. Many other concepts in mathematics, engineering, and physics are formulated as higher-order functions: convolution, filters, edge detectors, Fourier transforms, differential equations, Hamiltonians, etc. Even more sophisticated sorts of numerical computations that are difficult to express without the machinery of functional-programming languages, such as pumping methods for increasing rates of convergence, are persuasively discussed elsewhere [3] but stray beyond our present topic. If we are to raise the level of expressiveness of scientific programming we might wish to consider using similar conventions when coding such concepts. As we see below, with appropriate compilation technology, this can result in an *increase* in performance.

3 The AD Transforms *Are* Higher-Order Functions

The first argument `f` to the `nint` procedure of the previous section obeys a particular API: `nint` can call `f`, but (at least in any mainstream language) there are no other operations (with the possible exception of a conservative test for equality) that can be performed on a function passed as an argument. We might imagine improving `nint`'s accuracy and efficiency by having it use derivative information, so that it could more accurately and efficiently adapt its points of evaluation to the local curvature of `f`. Of course, we would want an AD transform of `f` rather than some poor numerical approximation to the desired derivative. Upon deciding to do this, we would have two alternatives. One would be to change the signature of `nint` so that it takes an additional argument `df` that calculates the derivative of `f` at a point. This alternative requires rewriting every call to `nint` to pass this extra argument. Some call sites would be passing a function argument to `nint` that is itself a parameter to the calling routine, resulting in a ripple effect of augmentation of various APIs. This can be seen above, where `nint2` would need to accept an extra parameter—or perhaps two extra parameters. This alternative, which we might call *caller-derives*, requires potentially global changes in order to change a local decision about how a particular numerical integration routine operates, and is therefore a severe violation of the principles of modularity.

The other alternative would be for `nint` to be able to internally find the derivative of `f`, in a *callee-derives* discipline. In order to do this, it would need to be able to invoke AD upon that function argument. To be concrete, we posit two derivative-taking operators which perform the forward- and reverse-mode AD transforms on the functions they are passed.[3] These have a somewhat complex API, so as to avoid repeated calculation of the primal function during derivative calculation. For forward-mode AD, we introduce $\overrightarrow{\mathscr{J}}$ which we for now give a simplified signature $\overrightarrow{\mathscr{J}} : (\mathbb{R}^n \to \mathbb{R}^m) \to ((\mathbb{R}^n \times \mathbb{R}^n) \to (\mathbb{R}^m \times \mathbb{R}^m))$. This takes a numeric function $\mathbb{R}^n \to \mathbb{R}^m$ and returns an augmented function which takes what the original function took along with a perturbation direction in its input space, and returns what the original function returned along with a perturbation direction in its output space. This mapping from an input perturbation to an output perturbation is equivalent to multiplication by the Jacobian. Its reverse-mode AD sibling has a slightly more complex API, which we can caricature as $\overleftarrow{\mathscr{J}} : (\mathbb{R}^n \to \mathbb{R}^m) \to (\mathbb{R}^n \to (\mathbb{R}^m \times (\mathbb{R}^m \to \mathbb{R}^n)))$. This takes a numeric function $\mathbb{R}^n \to \mathbb{R}^m$ and returns an augmented function which takes what the original function took and returns what the original function returned paired with a "reverse phase" function that maps a sensitivity in the output space back to a sensitivity in the input space. This mapping of an output sensitivity to an input sensitivity is equivalent to multiplication by the transpose of the Jacobian.

These AD operators are (however implemented, and whether confined to a pre-processor or supported as dynamic run-time constructs) higher-order functions, but they cannot be written in the conventional λ-calculus. The machinery to allow them to be expressed is somewhat involved [6, 7, 8].

Part of the reason for this complexity can be seen in `nint2` above, which illustrates the need to handle not only anonymous functions but also higher-order functions, nesting, and interactions between variables of various scopes that correspond to the distinct nested invocations of the AD operators. If `nint` is modified to take the derivative of its function argument, then the outer call to `nint` inside `nint2` will take the derivative of an unnamed function which internally invokes `nint`. Since this inner `nint` also invokes the derivative operator, the $\overrightarrow{\mathscr{J}}$ and $\overleftarrow{\mathscr{J}}$ operators must both be able to be applied to functions that internally

[3] One can imagine hybrid operators; we leave that for the future.

invoke $\overrightarrow{\mathscr{J}}$ and $\overleftarrow{\mathscr{J}}$. We also do not wish to introduce a new special "tape" data type onto which computation flow graphs are recorded, as this would both increase the number of data types present in the system, and render the system less amenable to standard optimizations.

Of course, nesting of AD operators is only one sort of interaction between constructs, in this case between two AD constructs. We wish to make all interaction between all available constructs both correct and robust. Our means to that end are uniformity and generality, and we therefore generalize the AD operators $\overrightarrow{\mathscr{J}}$ and $\overleftarrow{\mathscr{J}}$ to apply not only to numeric functions $\mathbb{R}^n \to \mathbb{R}^m$ but to any function $\alpha \to \beta$, where α and β are arbitrary types. Note that α and β might in fact be function types, so we will be assigning a meaning to "the forward derivative of the higher-order function `map`," or to the derivative of `nint`. This generalization will allow us to mechanically transform the code bodies of functions without regard to the types of the functions called within those code bodies. But in order to understand this generalization, we briefly digress into a mathematical domain that can be used to define and link forward- and reverse-mode AD.

4 AD and Differential Geometry

We now use some concepts from differential geometry to motivate and roughly explain the types and relationships in our AD-augmented λ-calculus. It is important to note that this is a cartoon sketch, with many details suppressed or even altered for brevity, clarity, and intuition.

In differential geometry, a differentiable manifold $\mathscr{N}$ has some structure associated with it. Each point $x \in \mathscr{N}$ has an associated vector space called its tangent space, whose members can be thought of as directions in which x can be locally perturbed in $\mathscr{N}$. We call this a *tangent vector* of x and write it $\overrightarrow{x}$. An element x paired with an element $\overrightarrow{x}$ of the tangent space of x is called a tangent bundle, written $\overrightarrow{x} = (x, \overrightarrow{x})$. A function between two differentiable manifolds, $f : \mathscr{N} \to \mathscr{M}$, which is differentiable at x, mapping it to $y = f(x)$, can be lifted to map *tangent bundles*. In differential geometry this is called the pushforward of f. We will write $\overrightarrow{y} = (y, \overrightarrow{y}) = \overrightarrow{f}(\overrightarrow{x}) = \overrightarrow{f}(x, \overrightarrow{x})$. (This notation differs from the usual notation of $T\mathscr{M}_x$ for the tangent space of $x \in \mathscr{M}$.)

We import this machinery of the pushforward, but reinterpret it quite concretely. When f is a function represented in a concrete expression in our augmented λ-calculus, we mechanically transform it into $\overrightarrow{f} = \overrightarrow{\mathscr{J}}(f)$. Moreover when x is a particular value, with a particular shape, we define the shape of $\overrightarrow{x}$, an element of the tangent space of x, in terms of the shape of x. If $x : \alpha$, meaning that x has type (or shape) α, we say that $\overrightarrow{x} : \overrightarrow{\alpha}$ and $\overrightarrow{x} : \overrightarrow{\alpha}$. These proceed by cases, and (with some simplification here for expository purposes) we can say that a perturbation of a real is real, $\overrightarrow{\mathbb{R}} = \mathbb{R}$; the perturbation of a pair is a pair of perturbations, $\overrightarrow{\alpha \times \beta} = \overrightarrow{\alpha} \times \overrightarrow{\beta}$, and the perturbation of a discrete value contains no information, so $\overrightarrow{\alpha} =$ **void** when α is a discrete type like **bool** or **int**. This leaves the most interesting: $\overrightarrow{\alpha \to \beta}$, the perturbation of a function. This is well defined in differential geometry, which would give $\overrightarrow{\alpha \to \beta} = \overrightarrow{\alpha} \to \overrightarrow{\beta}$, but we have an extra complication. We must regard a mapping $f : \alpha \to \beta$ as depending not only on the input value, but also on the value of any free variables that occur in the definition of f. Roughly speaking then, if γ is the type of the combination of all the free variables of the mapping under consideration, which we write as $f : \alpha \xrightarrow{\gamma} \beta$, then $\overrightarrow{\alpha \xrightarrow{\gamma} \beta} = \overrightarrow{\alpha} \xrightarrow{\overrightarrow{\gamma}} \overrightarrow{\beta}$. However we never map such raw tangent values, but always tangent bun-

dles. These have similar signatures, but with tangents always associated with the value whose tangent space they are elements of.

The powerful intuition we now bring from differential geometry is that just as the above allows us to extend the notion of the forward-mode AD transform to arbitrary objects by regarding it as a pushforward of a function defined using the λ-calculus, we can use the notion of a pullback to see how analogous notions can be defined for reverse-mode AD. In essence, we use the definition of a cotangent space to relate the signatures of "sensitivities" (our term for what are called adjoint values in physics or elements of a cotangent space in differential geometry) to the signatures of perturbations. Similarly, the reverse transform of a function is defined using the definition of the pullback from differential geometry.

If $\overrightarrow{f} : (x, \overrightarrow{x}) \mapsto (y, \overrightarrow{y})$ is a pushforward of $f : x \mapsto y$, then the pullback is $\overleftarrow{f} : \overleftarrow{y} \mapsto \overleftarrow{x}$, which must obey the relation $\overleftarrow{y} \bullet \overrightarrow{y} = \overleftarrow{x} \bullet \overrightarrow{x}$, where $\bullet$ is a generalized dot-product. If $\overrightarrow{\mathscr{J}} : f \mapsto \overrightarrow{f}$, then $\overleftarrow{\mathscr{J}} : f \mapsto (\lambda x \,.\, (f(x), \overleftarrow{f}))$, and some type simplifications occur. The most important of these is that we can generalize $\overrightarrow{\mathscr{J}}$ and $\overleftarrow{\mathscr{J}}$ to apply not just to *functions* that map between objects of any type, but to apply to *any* object of any type, with functions being a special case: $\overrightarrow{\mathscr{J}} : \alpha \to \overrightarrow{\alpha}$ and $\overleftarrow{\mathscr{J}} : \alpha \to \overleftarrow{\alpha}$. A detailed exposition of this augmented λ-calculus is beyond our scope here. Its definition is a delicate dance, as the new mechanisms must be sufficiently powerful to implement the AD operators, but not so powerful as to preclude their own transformation by AD or by standard λ-calculus reductions. We can however give a bit of a flavor: constructs like $\overrightarrow{\mathscr{J}}(\overleftarrow{\mathscr{J}})$ and its cousins, which arise naturally whenever there is nested application of the AD machinery, require novel operators like $\overleftarrow{\mathscr{J}}^{-1}$.

5 Migration to Compile Time

In the above exposition, the AD transforms are presented as first-class functions that operate on an even footing with other first-class functions in the system, like `+`. However, compilers are able to migrate many operations that appear to be done at run time to compile time. For instance, the code fragment `(2+3)` might seem to require a run-time addition, but a sufficiently powerful compiler is able to migrate this addition to compile time. A compiler has been constructed, based on the above constructs and ideas, which is able to migrate almost all scaffolding supporting the raw numerical computation to compile time. In essence, a language called VLAD consisting of the above AD mechanisms in addition to a suite of numeric primitives is defined. A compiler for VLAD called STALINGRAD has been constructed which uses polyvariant union-free flow analysis [10]. This analysis, for many example programs we have written, allows all scaffolding and function manipulation to be migrated to compile time, leaving for run time a mix of machine instructions whose floating-point density compares favorably to that of code emitted by highly tuned AD systems based on preprocessors and FORTRAN. Although this aggressive compiler currently handles only the forward-mode AD transform, an associated VLAD interpreter handles both the forward- and reverse-mode AD constructs with full general nesting. The compiler is being extended to similarly optimize reverse-mode AD, and no significant barriers in this endeavor are anticipated.

Although it is not a production-quality compiler (it is slow, cannot handle large examples, does not support arrays or other update-in-place data structures, and is in general unsuitable for end users) remedying its deficiencies and building a production-quality compiler would be straightforward, involving only known methods [5, 11]. The compiler's limitation to union-free analyses and finite unrolling of recursive data structures could also be relaxed using standard implementation techniques.

6 Some Preliminary Performance Results

We illustrate the power of our techniques with two examples. These were chosen to illustrate a hierarchy of mathematical abstractions built on a higher-order gradient operator [8]. They were *not* chosen to give an advantage to the present system or to compromise performance of other systems. They do however show how awkward it can be to express these concepts in other systems, even overloading-based systems.

Figure 1 gives the essence of the two examples. It starts with code shared between these examples: `multivariate-argmin` implements a multivariate optimizer using adaptive naïve gradient descent. This iterates $\mathbf{x}_{i+1} = \mathbf{x}_i - \eta \nabla f(\mathbf{x}_i)$ until either $\|\nabla f(\mathbf{x})\|$ or $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|$ is small, increasing η when progress is made and decreasing η when no progress is made. The VLAD primitives `bundle` and `tangent` construct and access tangent bundles, `j*` is $\overrightarrow{\mathscr{J}}$, and `real` shields a value from the optimizer. Omitted are definitions for standard SCHEME primitives and the functions `sqr` that squares its argument, `map-n` that maps a function over the list `(0` $\ldots n-1$`)`, `reduce` that folds a binary function with a specified identity over a list, `v+` and `v-` that perform vector addition and subtraction, `k*v` that multiplies a vector by a scalar, `magnitude` that computes the magnitude of a vector, `distance` that computes the l^2 norm of the difference of two vectors, and `e` that returns the i-th basis vector of dimension n.

The first example, `saddle`, computes a saddle point: $\min_{(x_1,y_1)} \max_{(x_2,y_2)} f(x,y)$ where we use the trivial function $f(x,y) = (x_1^2 + y_1^2) - (x_2^2 + y_2^2)$. The second example, `particle`, models a charged particle traveling non-relativistically in a plane with position $\mathbf{x}(t)$ and velocity $\dot{\mathbf{x}}(t)$ and accelerated by an electric field formed by a pair of repulsive bodies, $p(\mathbf{x};w) = \|\mathbf{x} - (10, 10 - w)\|^{-1} + \|\mathbf{x} - (10, 0)\|^{-1}$, where w is a modifiable control parameter of the system, and hits the x-axis at position $\mathbf{x}(t_f)$. We optimize w so as to minimize

```
(define ((gradient f) x)
 (let ((n (length x))) ((map-n (lambda (i) (tangent ((j* f) (bundle x (e i n)))))) n)))

(define (multivariate-argmin f x)
 (let ((g (gradient f)))
  (letrec ((loop (lambda (x fx gx eta i)
                   (cond ((<= (magnitude gx) (real 1e-5)) x)
                         ((= i (real 10)) (loop x fx gx (* (real 2) eta) (real 0)))
                         (else (let ((x-prime (v- x (k*v eta gx))))
                                 (if (<= (distance x x-prime) (real 1e-5))
                                     x
                                     (let ((fx-prime (f x-prime)))
                                      (if (< fx-prime fx)
                                          (loop x-prime fx-prime (g x-prime) eta (+ i 1))
                                          (loop x fx gx (/ eta (real 2)) (real 0))))))))))))
   (loop x (f x) (g x) (real 1e-5) (real 0)))))

(define (multivariate-argmax f x) (multivariate-argmin (lambda (x) (- (real 0) (f x))) x))

(define (multivariate-max f x) (f (multivariate-argmax f x)))

(define (saddle)
 (let* ((start (list (real 1) (real 1)))
        (f (lambda (x1 y1 x2 y2) (- (+ (sqr x1) (sqr y1)) (+ (sqr x2) (sqr y2)))))
        ((list x1* y1*) (multivariate-argmin
                         (lambda ((list x1 y1)) (multivariate-max
                                                 (lambda ((list x2 y2)) (f x1 y1 x2 y2)) start)) start))
        ((list x2* y2*) (multivariate-argmax (lambda ((list x2 y2)) (f x1* y1* x2 y2)) start)))
  (list (list (write x1*) (write y1*)) (list (write x2*) (write y2*)))))

(define (naive-euler w)
 (let* ((charges (list (list (real 10) (- (real 10) w)) (list (real 10) (real 0))))
        (x-initial (list (real 0) (real 8)))
        (xdot-initial (list (real 0.75) (real 0)))
        (delta-t (real 1e-1))
        (p (lambda (x) ((reduce + (real 0)) ((map (lambda (c) (/ (real 1) (distance x c)))) charges)))))
  (letrec ((loop (lambda (x xdot)
                   (let* ((xddot (k*v (real -1) ((gradient p) x))) (x-new (v+ x (k*v delta-t xdot))))
                    (if (positive? (list-ref x-new 1))
                        (loop x-new (v+ xdot (k*v delta-t xddot)))
                        (let* ((delta-t-f (/ (- (real 0) (list-ref x 1)) (list-ref xdot 1)))
                               (x-t-f (v+ x (k*v delta-t-f xdot))))
                         (sqr (list-ref x-t-f 0))))))))
   (loop x-initial xdot-initial))))

(define (particle)
 (let* ((w0 (real 0)) ((list w*) (multivariate-argmin (lambda ((list w)) (naive-euler w)) (list w0))))
  (write w*)))
```

Fig. 1. The essence of the `saddle` and `particle` examples.

Table 1. Run times of our examples normalized relative to a unit run time for STALINGRAD.

	Language/Implementation			
Example	STALINGRAD	ADIFOR	TAPENADE	FADBAD++
`saddle`	1.00	0.49	0.72	5.93
`particle`	1.00	0.85	1.76	32.09

$E(w) = x_0(t_f)^2$, with the goal of finding a value for w that causes the particle's path to intersect the origin.

Naïve Euler ODE integration ($\ddot{\mathbf{x}}(t) = -\nabla_{\mathbf{x}} p(\mathbf{x})|_{\mathbf{x}=\mathbf{x}(t)}$; $\dot{\mathbf{x}}(t+\Delta t) = \dot{\mathbf{x}}(t) + \Delta t\, \ddot{\mathbf{x}}(t)$; $\mathbf{x}(t+\Delta t) = \mathbf{x}(t) + \Delta t\, \dot{\mathbf{x}}(t)$) is used to compute the particle's path, with a linear interpolation to find the x-axis intersect (when $x_1(t+\Delta t) \le 0$ we let $\Delta t_f = -x_1(t)/\dot{x}_1(t)$; $t_f = t + \Delta t_f$; $\mathbf{x}(t_f) = \mathbf{x}(t) + \Delta t_f \dot{\mathbf{x}}(t)$ and calculate the final error as $E(w) = x_0(t_f)^2$.) The final error is minimized with respect to w by `multivariate-argmin`.

Each task models a class of real-world problems (rational agent-agent interaction and agent-world interaction) that appear in game theory, economics, machine learning, automatic control theory, theoretical neurobiology, and design optimization. Each also requires nesting: a single invocation of even higher-order AD is insufficient. Furthermore, they use standard vector arithmetic which, without our techniques, would require allocation and reclamation of new vector objects whose size might be unknown at compile time, and access to the components of such vectors would require indirection. They also use higher-order functions: ones like `map-n` and `reduce`, that are familiar to the functional-programming community, and ones like `gradient` and `multivariate-argmin`, that are familiar to numerical programmers. Without our techniques, these would require closures and indirect function calls to unspecified targets.

STALINGRAD performed a polyvariant union-free flow analysis on both of these examples, and generated Fortran-like code. Variants of these examples were also coded in SCHEME, ML, HASKELL, C++, and FORTRAN, and run with a variety of compilers and AD implementations. Here we discuss only the C++ and FORTRAN versions. For C++, the FADBAD++ implementation of forward AD was used, compiled with G++. For FORTRAN, the ADIFOR and TAPENADE implementations of forward AD were used, compiled with G77. In all variants attempts were made to be faithful to both the generality of the mathematical concepts represented in the examples and to the standard coding style of each language. This means in particular that "tangent-vector" mode was used where available, which put STALINGRAD at a disadvantage of about a factor of two. (Although STALINGRAD does not implement a tangent-vector mode it would be straightforward to add such a facility by generalizing `bundle` and `tangent` to accept and return lists of tangent values, respectively.)

Although the most prominent high-performance AD systems (ADIFOR, TAPENADE, and ADIC) claim to support nested use of AD operators, it is "well known" within the AD community they do not (Jean Utke, personal communication), as the present authors discovered when attempting to assess the performance of other AD systems on the above tasks. Implementing these examples in those systems required enormous effort, to diagnose the various warning and silently incorrect results and to craft intricate work-arounds where possible. These included both rewriting input source code to meet a variety of unspecified, undocumented, and unchecked restrictions, and modifying the output code produced by some of the tools [9]. Table 1 summarizes the run times, normalized relative to a unit run time for STALIN-

GRAD. Source code for all variants of our examples, the scripts used to produce Table 1, and the log produced by running those scripts are available at `http://www.bcl.hamilton.ie/~qobi/ad2008/`. This research prototype exhibits an increase in performance of one to three orders of magnitude when compared with the overloading-based forward AD implementations for both functional and imperative languages (of which only the fastest is shown) and roughly matches the performance of the transformation-based forward AD implementations for imperative languages.

7 Discussion and Conclusion

The TAPENADE 2.1 User's Guide [2, pp 72] states:

> **10. KNOWN PROBLEMS AND DEVELOPMENTS TO COME**
> We conclude this user's guide of TAPENADE by a quick description of known problems, and how we plan to address them in the next releases. [...] we focus on missing functionalities. [...]
> **10.4 Pointers and dynamic allocation**
> Full AD on FORTRAN95 supposes pointer analysis, and an extension of the AD models on programs that use dynamic allocation. This is not done yet.
> Whereas the tangent mode does not pose major problems for programs with pointers and allocation, there are problems in the reverse mode. For example, how should we handle a memory deallocation in the reverse mode? During the reverse sweep, the memory must be reallocated somehow, and the pointers must point back into this reallocated memory. Finding the more efficient way to handle this is still an open problem.

The Future Plans section on the OPENAD web site
`http://www-unix.mcs.anl.gov/~utke/OpenAD/` states:

> **4. Language-coverage and library handling in adjoint code**
> **2. language concepts (e.g., array arithmetic, pointers and dynamic memory allocation, polymorphism):**
> Many language concepts, in particular those found in object-oriented languages, have never been considered in the context of automatic adjoint code generation. We are aware of several hard theoretical and technical problems that need to be considered in this context. Without an answer to these open questions the correctness of the adjoint code cannot be guaranteed.

In PLT, semantics are defined by reductions which transform a program from the source language into the λ-calculus, or an equivalent formalism like SSA. Since we have defined the AD operators in a λ-calculus setting in an extremely general fashion, these operators interoperate correctly with all other constructs in the language. This addresses, in particular, all the above issues, and in fact all such issues: by operating in this framework, the AD constructs can be made available in a dynamic fashion, with extreme generality and uniformity. This framework has another benefit: compiler optimizations and other compiler and implementation techniques are already formulated in the same framework, which allows the AD constructs to be integrated into compilers and combined with aggressive optimization. This gives the numerical programmer the best of both worlds: the ability to write confidently in an expressive higher-order modular dynamic style while obtaining competitive numerical performance.

The λ-calculus approach also opens some exciting theoretical questions. The current system is based on the untyped λ-calculus. Can the $\overrightarrow{\mathscr{J}}$ and $\overleftarrow{\mathscr{J}}$ operators be incorporated into a typed λ-calculus? Many models of real computation have been developed; can this system be formalized in that sense? Can the AD operators as defined be proved correct, in the sense of matching a formal specification written in terms of limits or non-intuitive differential geometric constructions? Is there a relationship between this augmented λ-calculus and synthetic differential geometry? Could entire AD systems be built and formally proven correct?

Acknowledgement. This work was supported, in part, by NSF grant CCF-0438806, Science Foundation Ireland grant 00/PI.1/C067, and a grant from the Higher Education Authority of Ireland. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

1. Church, A.: The Calculi of Lambda Conversion. Princeton University Press, Princeton, NJ (1941)
2. Hascoët, L., Pascual, V.: TAPENADE 2.1 user's guide. Rapport technique 300, INRIA, Sophia Antipolis (2004). URL `http://www.inria.fr/rrrt/rt-0300.html`
3. Hughes, J.: Why functional programming matters. The Computer Journal **32**(2), 98–107 (1989). URL `http://www.md.chalmers.se/~rjmh/Papers/whyfp.html`
4. Johnsson, T.: Lambda lifting: Transforming programs to recursive equations. In: Functional Programming Languages and Computer Architecture. Springer-Verlag, Nancy, France (1985)
5. Nielson, F., Nielson, H.R., Hankin, C.: Principles of Program Analysis. Springer-Verlag, New York (1999)
6. Pearlmutter, B.A., Siskind, J.M.: Reverse-mode AD in a functional framework: Lambda the ultimate backpropagator. ACM Trans. on Programming Languages and Systems (2008). In press
7. Siskind, J.M., Pearlmutter, B.A.: First-class nonstandard interpretations by opening closures. In: Proceedings of the 2007 Symposium on Principles of Programming Languages, pp. 71–6. Nice, France (2007)
8. Siskind, J.M., Pearlmutter, B.A.: Nesting forward-mode AD in a functional framework. Higher-Order and Symbolic Computation (2008). To appear
9. Siskind, J.M., Pearlmutter, B.A.: Putting the automatic back into AD: Part I, What's wrong. Tech. Rep. TR-ECE-08-02, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA (2008). URL `ftp://ftp.ecn.purdue.edu/qobi/TR-ECE-08-02.pdf`
10. Siskind, J.M., Pearlmutter, B.A.: Using polyvariant union-free flow analysis to compile a higher-order functional-programming language with a first-class derivative operator to efficient Fortran-like code. Tech. Rep. TR-ECE-08-01, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA (2008). URL `http://docs.lib.purdue.edu/ecetr/367/`
11. Wadler, P.L.: Comprehending monads. In: Proceedings of the 1990 ACM Conference on LISP and Functional Programming, pp. 61–78. Nice, France (1990)