AD for an Array Language with Nested Parallelism

Schenck, Robert; Rønning, Ola; Henriksen, Troels; Oancea, Cosmin E.

Published in:
Proceedings of SC 2022

DOI:
10.1109/SC41404.2022.00063

Publication date:
2022

Document version
Peer reviewed version

Document license:
CC BY-SA

Citation for published version (APA):
AD for an Array Language with Nested Parallelism

Robert Schenck, Ola Rønning, Troels Henriksen and Cosmin E. Oancea
DIKU, University of Copenhagen, Denmark
r@bert.lol, ola@di.ku.dk, athas@sigkill.dk, cosmin.oancea@diku.dk

Abstract

We present a technique for applying (forward and) reverse-mode automatic differentiation (AD) on a non-recursive second-order functional array language that supports nested parallelism and is primarily aimed at efficient GPU execution.

The key idea is to eliminate the need for a "tape" by relying on redundant execution to bring into each new scope all program variables that may be needed by the differentiated code. Efficient execution is enabled by the observation that perfectly-nested scopes do not introduce re-execution, and such perfect nests are produced by known compiler transformations, e.g., flattening. Our technique differentiates loops and bulk-parallel operators—such as map, reduce, histogram, scan, scatter—by specific rewrite rules, and aggressively optimizes the resulting nested-parallel code. We report an experimental evaluation that compares with established AD solutions and demonstrates competitive performance on nine common benchmarks from recent applied AD literature.

Keywords: automatic differentiation, functional data parallel language, compilers, GPGPU.

1 Introduction

Automatic differentiation (AD) is a practical way for computing derivatives of functions that are expressed as programs. AD of sequential code is implemented in tools such as Tapenade [2], ADOL-C [14], and Stalingrad [35]. Modern deep learning is built on array programming frameworks such as Tensorflow [1] and PyTorch [32], which provide implicitly parallel bulk operations that support AD.

A largely-unsolved challenge is supporting AD for high-level parallel languages [18, 33, 44] that permit arbitrary nesting of sequential and parallel constructs. Such solutions may in principle act as a catalyst for prototyping and training of more advanced machine learning (ML) models.

ML involves minimising a cost function, which typically has far more inputs than outputs. The reverse mode of AD is the most efficient in such cases [3], but is challenging to implement because intermediate program values are required by the differentiated code. The program must first run a forward sweep (a.k.a., primal trace) that stores intermediate program states on the tape, which is then read during the return sweep, which essentially runs the program in reverse, and actually computes the derivative.

A significant amount of work has studied how to elegantly model reverse mode AD as a compiler transformation, and how to hide the tape under powerful programming abstractions such as (dynamic) closures [35] and delimited continuations [46]. These abstractions are not suited for efficient parallel execution on manycore hardware such as GPUs.

This work is to our knowledge the first to demonstrate an efficient GPU implementation of reverse mode AD as a compiler transformation on a data-parallel language that supports nested parallelism by means of higher-order array combinators (SOAC), such as map, reduce(-by-index), scan.

A key difference is that related approaches save by default all variables on the tape and support checkpointing annotations as an optimization (of memory footprint). However, in a nested-parallel context, the tape may give rise to complex, irregular data-structures that are passed across deep nests and that are challenging to implement efficiently in regards to optimizing spatial (coalescing) and temporal locality.

In contrast, our modeling of the tape takes inspiration from the idea [3] that applying reverse AD to a straight line of side-effect-free code does not require any tape per se, because all intermediate values remain available. We expand this idea to drive the code- transformation across lexical scopes by requiring that whenever the return sweep enters a new scope s, it first redundantly re-executes the forward sweep of s in order to bring all the needed variables into scope.

Our technique preserves the work and span asymptotics because the recomputation overhead is at worst proportional to the depth of the deepest nest of scopes, which is constant for a given nonrecursive program. Moreover, perfectly-nested scopes (other than loops) are guaranteed to not introduce re-execution, hence the overhead can be minimized by classic compiler transformations such as flattening nested parallelism [19] and polyhedral-like optimizations [7]. Since our tape is essentially formed by the in-scope variables, scalars are efficiently accessed from registers rather than global memory (tape), and the code resulted from AD fully benefits from the existent compiler- optimization repertoire.

In what sequential loops are concerned, the loop-variant variables require checkpointing because the return sweep needs to execute the iterations in reverse order. In addition, we exploit to some extent the fundamental time-space trade-off studied by Siskind and Pearlmutter [42], in a simple and practical way, by allowing the user to annotate how many times a loop should be strip-mined.\footnote{Strip-mining k times a loop of count $n^k$ requires only $2kn \times$ more memory than the original program at the expense of $k \times$ re-execution at worst. Of course, since the strip-mining factor is a user-defined constant it cannot achieve the logarithmic time-space overhead.}

A key difference is that related approaches save by default all variables on the tape and support checkpointing annotations as an optimization (of memory footprint). However, in a nested-parallel context, the tape may give rise to complex, irregular data-structures that are passed across deep nests and that are challenging to implement efficiently in regards to optimizing spatial (coalescing) and temporal locality.
Having designed the glue that binds scopes together, we turn our attention to deriving high-level rewrite rules for differentiating the parallel operators of the language. We achieve this by starting from the main rewrite rule of the reverse mode\(^2\) and extend it by applying reasoning that combines imperative (dependence analysis, loop distribution) and functional thinking (rewrite rules, recurrences as scans).

In particular, the simplest parallel operator, namely map, is the most difficult one to translate, because its purely functional semantics allows free variables to be freely read inside it, but reverse AD replaces a read with an accumulation, which cannot be in general represented as a combination of classical data-parallel constructs. In this context we report optimizations related to turning accumulators to more-specialized constructs, such as reductions and generalized histograms [17], that can yield speedups close to one order of magnitude at the application level (e.g., GMM and LSTM).

Our overall contribution is an end-to-end algorithm of AD with support for nested parallel combiners, implemented inside (a clone of) the compiler for the Futhark language and aimed at GPU execution. Our specific contributions are:

- A redundant-execution technique for reverse AD that (i) eliminates the need for tape, (ii) is guaranteed to not introduce re-execution for perfectly-nested scopes other than loops, and (iii) practically supports the time-space tradeoff by means of loop stripmining.
- A set of rewrite rules for differentiating higher-order parallel combiners, including uses of free variables.
- A collection of optimizations that rewrite common cases of accumulators to other specific language constructs that benefit from specialized code generation.
- An experimental evaluation that demonstrates (i) sequential performance competitive with Tapenade [2] on ADBench [43], (ii) GPU performance competitive with Enzyme [29] on two of their applications, i.e., RS-Bench, XSBench, and (iii) significant GPU speedups in comparison with PyTorch [32] on GMM, LSTM, and for the sparse formulation of K-Means.

2 Preliminaries

This section provides the gist of how automatic differentiation of programs may be derived from its mathematical foundation. Given a function \( P : \mathbb{R}^a \rightarrow \mathbb{R}^d \), the (total) derivative of \( P \) at a point \( x \in \mathbb{R}^a \) (where \( P \) is differentiable) can be written as a function of \( y \in \mathbb{R}^a \) in terms of its Jacobian as \( \frac{dP(x)}{dy}(y) = J_P(x) \cdot y \), where \( \cdot \) is matrix-vector multiplication.

The Jacobian is a representation of a function derivative at a specific point. The chain rule provides a straightforward way of deriving the derivative across function composition.

\[ \frac{dP(x)}{dx}(x) = \frac{dP(x)}{dy}(y) \cdot \frac{dy}{dx}(x) \]

\[ \frac{d}{dx}(f(x)) = f'(x) \cdot \frac{dx}{dx} \]

\[ \frac{d}{dx}(f(a, b)) = \frac{\partial f}{\partial a}(a, b) \cdot \frac{da}{dx} + \frac{\partial f}{\partial b}(a, b) \cdot \frac{db}{dx} \]

Figure 1. (a) A program \( P \) which computes the function \( f(x_0, x_1) = (x_1 \cdot \sin(x_0), x_0 \cdot x_1) \), (b) the forward mode AD transformation of the program, and (c) the reverse mode AD transformation of the program.

For example, if \( P(x) = h(g(f(x))) \), and the types of \( f, g, h \) are \( f : \mathbb{R}^a \rightarrow \mathbb{R}^b, g : \mathbb{R}^b \rightarrow \mathbb{R}^c \), and \( h : \mathbb{R}^c \rightarrow \mathbb{R}^d \), then:

\[ J_P(x) = J_h(g(f(x))) \cdot J_g(f(x)) \cdot J_f(x) \]

(1)

Since matrix multiplication is associative, there are multiple ways to evaluate \( J_P(x) \) when it’s viewed as a program. For example, from right-to-left:

\[ J_P(x) = J_h(g(f(x))) \cdot [J_g(f(x)) \cdot J_f(x)] \]

or left-to-right:

\[ J_P(x) = [J_h(g(f(x))) \cdot J_g(f(x))] \cdot J_f(x) \]

Notice that if \( h \) is a scalar function (i.e., \( d = 1 \)), then \( J_h(g(f(x))) \) is a gradient vector of dimension \( \mathbb{R}^d \). In the right-to-left formulation, \( \star \) is a large matrix of dimension \( \mathbb{R}^{C \times a} \), whereas in the left-to-right formulation, \( \star \) is a vector of dimension \( \mathbb{R}^c \). When \( d \ll a \), it is much more expensive to compute/manifest the larger intermediate Jacobian matrices in the right-to-left evaluation order. When \( a \ll d \), it’s more efficient to choose the right-to-left evaluation order. There’s a catch: the Jacobians depend on intermediate results, e.g., \( J_g(f(x)) \) depends on \( f(x) \). In the right-to-left evaluation order, the computation of these intermediate results can be interwoven with the computation of the Jacobians, as the Jacobians are computed in program-order. In the left-to-right evaluation there’s no such luck: \( P(x) \) must first be executed, with all intermediate results saved, before the \( J_P(x) \) may be computed. For this reason, when \( a \approx d \) the right-to-left evaluation order is preferred.

2.0.1 Forward mode AD. Consider the example program \( P \) in Figure 1a. The forward mode AD transformation of \( P \), \( \tilde{P} \), is shown in Figure 1b; the transformation involves the core rewrite rule:\(^3\)

\[ \text{let } v = f(a, b) \text{ implies let } \tilde{v} = \frac{\partial f(a, b)}{\partial a} \tilde{a} + \frac{\partial f(a, b)}{\partial b} \tilde{b} \]

\( \tilde{a} \) is an original-program statement \( x = f(a) \) produces the differentiated code \( \tilde{x} = \frac{\partial f(a)}{\partial a} \cdot \tilde{a} \). where \( \tilde{a} \) the adjoint of program variable \( x \).

\( \tilde{a} \) is an original-program statement \( x = f(a) \) produces the differentiated code

\( \tilde{x} = \frac{\partial f(a)}{\partial a} \cdot \tilde{a} \). where \( \tilde{a} \) the adjoint of program variable \( x \).

\( \tilde{a} \) is an original-program statement \( x = f(a) \) produces the differentiated code

\( \tilde{x} = \frac{\partial f(a)}{\partial a} \cdot \tilde{a} \). where \( \tilde{a} \) the adjoint of program variable \( x \).

This rule generalizes for \( n \)-array functions in the obvious way: for any statement \( \text{let } v = f(a_0, \ldots, a_{n-1}) \), we have \( \tilde{v} = \sum_{i=0}^{n-1} \frac{\partial f}{\partial a_i} \cdot \tilde{a}_i \).
We expose forward mode AD to the user via a function which computes \( \text{dotacc} \) with respect to a chosen direction \( \text{dotacc} \). This means that for any differentiable program \( P \), \( \text{dotacc} \) takes a point and a direction and returns the derivative of \( P \) at the point in the right-to-left association of the Jacobian for \( P \). In order to recover the full Jacobian \( \nabla P(x) \), \( \nabla P \) is mapped over the standard basis of \( \mathbb{R}^n \).

2.0.2 Reverse mode AD. The reverse mode AD transformation of \( P \), \( \nabla P \), is shown in Figure 1c. As with the forward mode transformation, it involves the core rewrite rule:

\[
\text{let } v = f(a,b) \\
\text{let } u = f(a,b) \implies \text{let } \overline{a} = \frac{\partial f(a,b)}{\partial \overline{a}} \\
\text{let } \overline{b} = \frac{\partial f(a,b)}{\partial \overline{b}}
\]

where \( \overline{u} \) is the adjoint of \( v \), which is the derivative of \( P \) with respect to \( u \). The adjoint of a variable \( a \) may be updated multiple times—any use of \( a \) may contribute to the derivative (see, e.g., \( x_1 \) in Figure 1c). Reverse mode AD corresponds to computing the left-to-right Jacobian \( \nabla P(x) \). All intermediate results of \( P \) which appear in adjoint expressions must first be saved, which corresponds to computing \( P \) before the adjoint of each program variable. The \( \overline{\cdot} \) indicates the presence of these statements, along with any preceding adjoint expressions. The adjoints in Figure 1c are computed in the reverse program order, corresponding to a left-to-right association of the Jacobian of \( P \). Any adjoint \( \overline{a} \) is necessarily finalized before it is used: all uses of \( a \) must occur after assignment; any contributions to \( \overline{a} \) must necessarily appear before any uses of \( \overline{a} \) in reverse program order. \( \overline{P} \) takes as additional arguments the initial adjoints of the output \( x_0 \) and \( y_1 \) and additionally returns the adjoints of the input, \( x_0 \) and \( x_1 \). Reverse mode AD is made available to the user via a function:

\[
\text{vjp} : (P : \mathbb{R}^n \rightarrow \mathbb{R}^m) \rightarrow (\nabla P : \mathbb{R}^n \rightarrow \mathbb{R}^m \rightarrow \mathbb{R}^n)
\]

which computes \( \nabla P \) for any differentiable program \( P \), with the property that \( (\nabla P(x))((\mathbf{y})) = \frac{\delta}{\delta \mathbf{y}} P(x) \). As \( \nabla P \) has the same computational properties as \( \nabla P(x) \), it is the preferred choice when \( m \ll n \). When \( P \) computes a scalar function (i.e., \( m = 1 \)), \( \text{vjp} \) returns the complete gradient of \( P \) in a single pass; \( \text{vjp} \) requires \( n \) passes, one for each basis vector of \( \mathbb{R}^n \). Programs with very high-dimensional inputs and low-dimensional outputs appear frequently in the fields of optimization and machine learning, making \( \text{vjp} \) the computationally superior choice. The \( \text{vjp} \) transformation is also more complex due to computing adjoints in reverse order, storing of intermediate program variables, and accumulation of derivatives into adjoint variables, which turn any read of a variable in the original program into a write in the transformed program. For these reasons, we concentrate our discussion on reverse mode AD.

Higher-order derivatives are supported by nesting of \( \text{vjp} \) or \( \text{vjp} \).

2.1 Language

We perform our transformation on a data-parallel language which features arbitrary nesting of second-order array combinators (SOACs), e.g., \text{map}, \text{reduce}, and \text{scan}. SOACs are variadic in their number of arguments and returns, i.e., \text{zips} and \text{unzips} are implicit. For example,

\[
\text{unzip3} \left( \lambda(a, b, c) \rightarrow (c, b, a) \right) \text{ (zip3 as bs cs)}
\]

may be equivalently written as simply

\[
\text{map} \left( \lambda a b c \rightarrow (c, b, a) \right) \text{ as bs cs}
\]

The source language supports higher-order functions, polymorphism, modules, and similar high-level features, which are compiled away using a variety of techniques [12, 20] before we perform AD. The only remaining second-order functions are the SOACs. Lambdas can only appear syntactically in SOACs and VJP, and are not values. As such we do not suffer from “perturbation confusion” [26]. Further, a significant battery of standard optimisations (CSE, constant folding, aggressive inlining) is also applied prior to AD.

The language is written in A-normal form [39] (ANF): all subexpressions are variables names or constants with the exception of the body expression of \text{loops}, \text{if}-expressions and \text{let}-expressions. \text{let}-expressions consist of a series of bindings—that we also call statements—followed by a sequence of one or more returns:

\[
\text{let } a = 5 \ast x \text{ let } b = a \ast a \text{ in } b
\]

The language is purely functional: re-declarations of the same variable (in a given scope) should be understood as a notational convenience for variable shadowing. It supports a functional flavor of in-place updates based on uniqueness types [18]. The binding \text{let } xs[i] = x \text{ is syntactic sugar for } \text{let } x_{s'} = xs \text{ with } [i] \leftarrow x, \text{ which has the semantics that } x_{s'} \text{ is a copy of } xs \text{ in which the element at index } i \text{ is updated to } x, \text{ but also provides the operational guarantee that the update will be realized in place.}

The language also features sequential and pure loops, which have the semantics of a tail-recursive function:

\[
\text{let } y = \text{loop} \ (x) = (x_0) \text{ for } i = 0 \ldots n - 1 \text{ do } e
\]

The loop is initialized by binding \( x_0 \) to \( x \). Each iteration of the loop executes \( e \), and binds the result of the expression to \( x \), which is used on the subsequent iteration of the loop.
We implement forward mode AD by straightforward applications of the forward mode rewrite rule (2). That is, tangent variables statements are interleaved with primal variable statements (tangent variables are associated with primal variables via a simple mapping). In parameters, body results, and SOAC arguments we bundle tangent variables together with their primal counterpart.

As an alternative to a program transformation, forward mode AD can be implemented by coupling primal and tangent variables together in a tuple, e.g., \((\alpha, \dot{\alpha}) = (f(a), \frac{df}{da})\). Via overloading, the operators in the program may operate directly on such tuples, giving rise to the operator-overloading/dual number formulation of forward-mode AD described in [3]. Our approach of primal and tangent statements and bundling of primal and tangent variables is equivalent to this formulation of AD, implemented as a compiler pass instead of via operator overloading.

4 Reverse-AD by Redundant Execution

We first demonstrate the gist of our technique on a perfect-nest example in Section 4.1, then present the code transformation across scopes in a semi-formal manner in Section 4.2, and conclude with an example demonstrating our code generation for loops and the time-space trade-off in Section 4.3.

4.1 The Gist of Our Technique

Figure 2 shows the code generated by applying reverse-mode AD to the function below, and to the result adjoint `xss`?

\[
\lambda \text{ass} \rightarrow \text{let } xss = \text{map}(\lambda c \rightarrow \text{if } c \text{ then } \ldots \text{ else } \text{map}(\lambda a \rightarrow a^2a) \text{ as } \text{cs ass }) \text{ in } xss
\]

The program has four scopes: the outermost one (scope 0), the one of the outer-map lambda (scope 1), the one of the then/else branches (scope 2), and the one of the innermost lambda (scope 3). While generating the return sweep, our technique re-executes the forward sweep of the corresponding scope (1, 2 and 3) in order to make available the original program variables that might be used in the differentiated code. Essentially, the forward sweep is executed at worst four times, i.e., the depth of the deepest program nest.

However, it is easy to observe that perfect nests (other than loops) are guaranteed to not introduce re-computation, because, by definition, their bodies consists of one (composed) statement, hence the scope result cannot possibly be used by the return sweep. In Figure 2 the bindings of `xss, xs, xs’ and x are all dead code, hence the differentiated program after optimization suffers no re-computation overheads:

\[
\text{let } xss = \text{map}(\lambda c \rightarrow \text{if } c \text{ then } \ldots \text{ else } \text{map}(\lambda a \rightarrow a^2a) \text{ as } xss) \text{ in } xss
\]

It follows that overheads can be optimized by known transformations [7, 45] rooted in loop distribution and interchange, also supported by Futhark [19]. With their help, we commonly expect the forward sweep to be executed twice:

1 once for the outermost scope, as expected, because programs typically consist of multiple nests and also the user may require the result of the original program,

2 once for the innermost scope that typically performs scalar computation, which are cheap to recompute. In comparison, vectorization or the use of the tape would require such scalars to be retrieved from global memory, which is about two order of magnitude slower.

4.2 Transformation Rules Across Scopes

Figure 3 sketches the reverse-mode AD code transformation at the lexical levels of a body of statements \((\sigma P_{body})\), a sequence of statements \((\sigma P_{stmt})\) and individual statements \((\sigma P_{stmt})\). We recall that \(body\) refers to a sequence of \(let\) bindings (statements) followed by an \(in\) result. Our presentation sacrifices some formalism for readability, for example, by omitting the environment, which essentially carries out a mapping from the variables of the source program to their

\(*\) Map independently applies a function argument to each of the elements of an array (argument), i.e., \(\text{map } f [x_1, \ldots, x_n] = [f x_1, \ldots, f x_n]\).
The \( vjp_{\text{body}} \) rule highlights the redundant-execution mechanism that removes the need to implement the tape as a separate abstraction: each statement \( stm \) is processed individually (by \( vjp_{\text{stm}} \)), producing a sequence of statements on the forward sweep, denoted \( \overrightarrow{stm} \), that brings into scope whatever information is necessary to execute the return sweep for that statement, denoted by \( \overleftarrow{stm} \).

For example, the forward sweep of a multiplication statelet \( x = a \times b \) is the statement itself. Re-executing it brings into scope the (value of) variable \( x \) which may be needed in the return sweep of a following statement, e.g., let \( y = x \times a \) which would require an update to the adjoint of \( a \) according to the rewrite rule of Equation (3): \( \overrightarrow{a} \leftarrow \frac{\partial(xa)}{\partial a} \overrightarrow{y} \), which results in \( \overrightarrow{a} \leftarrow x \times \overrightarrow{y} \).

The rule for a loop statement is more complex. Figure 3 shows a loop that exhibits a loop-variant variable \( x \)—whose value changes through the loop—which is initialized with \( x_0 \) just before the loop is started, and whose value for the next iteration is bound to the result of the loop body. The result of the entire loop is bound to variable \( y \). The rule assumes that the shape of \( x \) does not change throughout the loop.

For example, the forward sweep of a multiplication statement let \( y = x \times a \) of a loop, except that its body is modified to checkpoint into array \( x \) the value of \( x \) at the entry of each iteration. As such \( x \) is also declared as loop variant and initialized to \( x_0 \), which, in turn, is allocated (by scratch) just before the loop start. Note that only the loops of the current scope are checkpointed; an innermost loop would be re-executed but not checkpointed.

The return sweep consists mainly of a loop that iterates with \( i \) from \( n-1 \) down to 0: the first statement re-installs the value of \( x \) for the current iteration from the checkpoint (i.e., \( x[i] \)). The remaining statements, \( \overleftarrow{stm} \), are those generated by \( vjp_{\text{body}} \) from the original loop body, including its forward sweep. As \( \overleftarrow{stm} \) may use \( x \), re-installment is necessary. Re-execution of the forward sweep brings into scope all variables that might possibly be used by the return sweep of the body.

The result of a reversed iteration is the adjoint of the original result \( \overrightarrow{r} \), together with the (updated) adjoints of the free variables used in the loop \( \overrightarrow{fstm} \). These are declared as variant through the loop \( \overrightarrow{fstm} \), such that the updates of all iterations are recorded. \( \overrightarrow{fstm} \) is the adjoint of the free variables prior to entering the loop; if a free variable was used in a prior statement of the same scope (and hence has an existing potentially non-zero adjoint) its adjoint is simply retrieved. Otherwise, its adjoint is initialized to a zero element of the appropriate type and shape.

---

8 The statements of the forward and return sweeps are arranged symmetrically; all statements of the forward sweep come before any of the return sweep, and the latter is organized in the reverse order of the original statements: \( vjp_{\text{stm}}(stm, \overrightarrow{stm}) \Rightarrow \overleftarrow{stm}, vjp_{\text{stm}}(stm, \overrightarrow{stm}) \Rightarrow \overleftarrow{stm} \).
When generating code for the return sweep of a statement scope for the source code shown in Listing 1, which again, similar to the in-place updates of Section 2.1. The final statement of the return sweep is returned. This is because the first (implicit) statement of the original loop is let y = a[i], hence it is differentiated last on the return sweep. Moreover, the loop returns the (updated) adjoint of its variant variable x, which is bound/saved in x′, and is thus used to update the adjoint of x₀.

The latter is necessary because otherwise, the need of checkpointing will keep them alive, and comes at the cost of (asymptotically) increasing the memory footprint. The time-space tradeoff [42] is demonstrated in Listing 2: applying the reverse AD to the strip-mined loop \( k = \log_{m} m^k \) times increases the re-execution factor from 2x \( \times (k + 2) \times \), but requires only \( mk \) more memory than the original program—i.e., the checkpointing of each of the \( k \) strip-mined loops stores \( m \) versions of the loop-variant variable \( y \); when \( m \) is constant, this results in logarithmic space and time overhead.

However, since GPUs have relatively small caches, we have found this tradeoff useful only when AD runs out of memory. In this sense, we allow the user to annotate the loops with a constant strip-mining factor, which is applied automatically before the application of reverse AD.

5 Rewrite Rules for Parallel Constructs

Having defined the transformation across scopes, this section focuses on the mapping of individual parallel constructs within a scope, though we allow arbitrary nesting of them and loops together. We organize the discussion in free-writing style fashion that focuses on the reasoning that led to the derivation of the rules rather than presenting them in a formal, but dry and verbose notation. Some of these rules resemble some that have been published previously [34], but we developed them independently [30].

5.1 Reduce and Multi-Reduce

We recall that the semantics of reduce, for an arbitrary binary associative operator \( \odot \) and its neutral element \( e_0 \) is:

\[
\text{reduce} \odot e_0 \ [a_0, a_1, \ldots, a_{n-1}] \equiv a_0 \odot a_1 \odot \ldots \odot a_{n-1}
\]

If the result of reduce is let-bound to variable \( y \), we can reason more easily about the contribution of the reduce statement to the adjoint of \( a_i \) if we group terms as:

\[
\text{let } y = (a_0 \odot \ldots \odot a_{i-1}) \odot a_i \odot (a_{i+1} \odot \ldots \odot a_{n-1})
\]

If we would know, for every \( i \), the terms \( l_i = a_0 \odot \ldots \odot a_{i-1} \) and \( r_j = a_{i+1} \odot \ldots \odot a_{n-1} \), we could directly apply the main rule for reverse AD given in Equation (3), which results in:

\[
\frac{\partial f}{\partial a_i} = \frac{\partial (l_i \odot a_i \odot r_j)}{\partial a_i} y
\]

where \( l_i \) and \( r_j \) are considered constants and \( \nabla \) denotes a potentially vectorized addition that matches the datatype.
The code for the right-hand side (RHS) can be generically generated as a function \( f \) that can be mapped to all \( a_i, l_i, r_i \):

\[
 f \leftarrow \forall j \in \lambda (l_i, a_i, r_i) \rightarrow l_i \odot a_i \odot r_i
\]

except that the adjoints of \( l_i \) and \( r_i \) are not returned by \( f \).

Finally, all the \( l_i \) and \( r_i \) values can be computed by a forward and reverse exclusive scan (prefix sum), respectively. Essentially, the forward sweep is the reduce statement. Denoting by \( as = [a_0, a_1, \ldots, a_{n-1}] \), and assuming for simplicity that \( \odot \) has no free variables, the return sweep is:

\[
\text{let } ls = \text{scan}^{exx} \odot e_\otimes \text{as}
\]

\[
\text{let } rs = \text{reverse as } \rightarrow \\
\text{scan}^{exx}(\lambda x y \rightarrow y \odot x) e_\otimes \rightarrow \text{reverse}
\]

\[
\text{let } \overline{as} = \text{map } f \text{ ls as rs}
\]

Essentially, the reverse AD code for an arbitrary reduce requires a map, two scan\(^{10}\) and two reverse operations, which preserve the work-depth parallel asymptotics of the original program. In practice, this is quite expensive,\(^{11}\) but luckily most standard operators, admit more efficient translations:

5.1.1 Special Cases of Reduce Operators. The treatment of plus, min, max is known [21], but we present a more efficient rule for multiplication. We reconfigure here for completeness and because they are used for reduce-by-index, which, to our knowledge, has not been covered before.

When the reduce operator is (vectorized) plus, we have

\[
\frac{\partial (l_i + a_i * r_i)}{\partial a_i} \bar{y} = \bar{y}, \text{ hence the return sweep becomes}
\]

\[
\text{let } \overline{as} = \bar{y}, \text{ which is also derived automatically by the}
\]

simplification engine from the general rule.

When the reduction operator is multiplication, we have

\[
\frac{\partial (l_i * a_i * r_i)}{\partial a_i} \bar{y} = l_i * r_i * \bar{y}. \text{ We discriminate three cases here:}
\]

- if all \( a_s \)'s elements are nonzeros, then \( l_i * r_i = \frac{\bar{y}}{a_i} \) and \( y \neq 0 \), hence we update each element: \( \overline{a_i} + \bar{x} = \frac{\bar{y}}{a_i} \).
- if exactly one element at index \( i_0 \) is zero, then \( l_i * r_i \) is zero for all other elements, and \( \overline{a_i} + \bar{x} = y * \bar{y} \).
- otherwise, \( \forall i, l_i * r_i \equiv 0 \), and \( \overline{as} \) remains unchanged.

The forward sweep is modified to compute the number of zeros in \( as \) and the product of non-zero elements (by a map-reduce operation), followed by setting the reduced result \( y \) accordingly. The return sweep computes the contributions by a parallel map and updates adjoints as discussed before.

Finally, when the reduction operator is \( \text{min} \) (or \( \text{max} \)), then only the adjoint of (one of) the minimal array element should receive the contribution of \( y \), because it is the only one that was used to compute the result. As such the forward sweep is modified to compute the minimal element together with its (first) index \( i_y \) (the common "argmin" operation), which can be done with a parallel reduction. The forward sweep is then \( \text{let } (y, i_y) = \text{argmin as} \), and the return sweep could be \( \overline{as}[i_y] = \bar{y} \).

However, at this point we add a compiler improvement that exploits sparsity: For example, if \( \overline{as} \) has not been created yet, then instead of inserting the update statement, we (statically) record in \( \forall j/p \)'s environment that \( \overline{as} \) is sparse (currently one non-zero element). Further operators such as reduce\(^{12}\) may refine the sparse structure of \( \overline{as} \) by adding more points to it. If \( as \) is created by a map operation, and \( \overline{as} \) is still sparse, then we (statically) replace the adjoint code of the map with (only) the adjoint code of the iterations corresponding to the non-zeros in \( \overline{as} \)—because the other iterations have no effect.

5.1.2 Reduce-by-Index. Reduce-by-index [17], a.k.a. multi-reduce, essentially generalizes a histogram computation by allowing the values from an array \( as \) that fall into the same bin (index from \( \text{inds} \)) to be reduced with an arbitrary associative and commutative operator \( \odot \) having neutral element \( e_\otimes \), where the number of bins \( m \) is typically assumed smaller than that of index-value pairs \( n \), i.e.,

\[
\text{let } hs = \text{reduce_by_index } (\odot) e_\otimes \text{inds as}
\]

has the semantics:

\[
\text{loop } hs = \text{replicate } m e_\otimes \text{inds as}
\]

\[
\text{for } i = 0 \ldots n-1 \text{ do}
\]

\[
\text{let } hs[\text{inds}[i]] \odot= as[i] \text{ in } hs
\]

A similar reasoning as for the arbitrary reduce suggests that two scans need to be applied to each subset of elements that fall in the same bin, a.k.a., multi-scan, in order to compute the \( l_i \) and \( r_i \) terms for every \( i \). Then the contributions to the adjoint of \( as \) of a reduce-by-index statement are computed as before by \( \text{map } f \) \( 1s \) as \( rs \). Assuming a constant key size, the multi-scan can be implemented within the right work-depth asymptotic by (radix) sorting \( as \) according to the corresponding bins, and then by applying irregular-segmented scans (forward and reverse) on the result. Work is in progress to implement this case. We have implemented however the special-case operators discussed for reduce. Essentially, the forward sweep consists of the reduce-by-index statement, but enhanced with the extended operators, and the return sweep is similar to reduce, except that in the update formula of the adjoint \( \overline{as}[i] \), \( \bar{y} \) is replaced with \( hs[\text{inds}[i]] \).

5.2 Scan or Prefix Sum

An inclusive scan [5] computes all prefixes of an array by means of an associative operator \( \odot \) with neutral element \( e_\otimes \):

\[
\text{scan} \odot e_\otimes [a_0, \ldots, a_{n-1}] \equiv [a_0, a_0 \odot a_1, \ldots, a_0 \odot \ldots \odot a_{n-2}]
\]

While the derivation of (multi-) reduce builds on a functional-like high-level reasoning, in scan’s case, we found it easier to reason in an imperative, low-level fashion. For simplicity we assume first that \( \odot \) operates on reals, and generalize later.
The loop above that implements scan, writes each element of the result array \( rs \) exactly once. To generate its return sweep, we can reason that we can fully unroll the loop, then apply the main rewrite-rule from Equation (3) to each statement and finally gather them back into the loop below:

\[
\begin{align*}
\text{let } & \quad \ell s(0) = a s[0] \\
\text{let } & \quad \ell s = \text{loop } (r s) = (r s) \text{ for } i = 1 \ldots n - 1 \text{ do} \\
\text{let } & \quad \ell s[i] = \ell s[i - 1] \odot \ell s[1] \text{ in } r s
\end{align*}
\]

It follows that in the rewrite rule above, \( \text{case when scan’s element type is one scalar. We also support vectorized such operators by turning them beforehand to a regular-segmented scan, by the rule:}

\[
\text{scan (map (○)) } y z \Rightarrow \text{transpose } x s \rightarrow \text{map (scan ⊲ e ○) } \rightarrow \text{transpose}
\]

5.3 Parallel Scatter

The statement \( \text{let } y s = \text{scatter } x s \) is vs produces an array \( y s \) by updating in-place the array \( x s \) (which is consumed) at the \( m \) indices specified in array \( y s \) with corresponding values taken from \( x s \). Scatter has constant depth, and work proportional with \( m \), the size of array \( y s \), i.e., it does not depend of the length \( n \) of the updated array \( y s \). Our rule assumes that is contains no duplicates, i.e., idempotent updates are currently not supported.

Denoting by \( \text{gather} \) the operation that reads from a support array elements at a given subset of indices, i.e.,

\[
\text{let } \quad \text{gather arr inds } = \text{map } (\lambda i \to x s[i]) \text{ inds}
\]

the forward sweep saves (in \( x s_{\text{s\text{aved}}} \)) the elements of \( x s \) that are about to be overwritten prior to performing the update:

\[
\text{let } x s_{\text{s\text{aved}}} = \text{gather } x s
\]

Since the original statement performs \( y s[i] = \text{vs}(i) \) for any \( i \in \text{inds} \), we have by Equation (3) \( y s[i] = y s[\text{inds}(i)] \). As such, the return sweep (1) first updates the adjoint of \( y s \), then (2) creates the adjoint of \( x s \) by zeroing out the elements from \( y s \) that were subject to the scattered update—because those adjoints correspond to elements that were never in \( x s \), and finally, (3) restores \( x s \) to its state before the update:

\[
\text{let } x s = \text{scatter } y s \Rightarrow x s_{\text{s\text{aved}}}
\]

Note that both forward and return sweep preserve the original work-depth asymptotics, because all operations have work proportional to \( m \) (not \( n \)). This would not hold if, e.g., the forward sweep would make a copy of the whole \( x s \).

5.4 Map

A \text{map} applies a lambda function to each element of an array, producing an array of same length:

\[
\text{let } x s = \text{map } (\lambda a \to \text{stms in } x) \text{ as}
\]
If the lambda has no free variables, the return sweep is:

$$\text{let } \overline{\alpha} = \text{map} (\lambda(a, \overline{a}, \overline{x}) \rightarrow \overline{stms} \downarrow \overline{stms} \in \overline{a}_0 + \overline{a})$$

as \( \overline{as} \rightarrow \overline{x} \overline{s} \)

where \( \overline{stms} \) and \( \overline{stms} \) denote the forward and reverse statements of lambda’s body, and \( \overline{a} \) is shadowed by \( \overline{stms} \).

A naive way of handling free variables is to turn them into bound variables. E.g. converting \( \text{map} (\lambda(i \rightarrow \overline{a}[i]) \rightarrow \overline{as}[i]) \) is into \( \text{map} (\lambda(i, \overline{a}') \rightarrow \overline{as}[i]) \) is (\( \text{replicate} \ n \overline{as} \)) where \( n \) is the size of \( \overline{is} \). This is fine for scalars, but asymptotically inefficient for arrays that are only partially used, as here, as the adjoin will be mostly zeroes.

In an impure language, we could update the adjoin of a free array variable \( \overline{as}[i] \) with an operation \( \overline{as}[i] = \overline{a} \), implemented with atomics or locks in the parallel case. In our pure setting, we instead introduce accumulators. An array can be “temporarily” turned into an accumulator with \( \text{withacc} \): 

$$\text{withacc} : [d] \alpha \rightarrow (\text{acc}(\alpha) \rightarrow \text{acc}(\alpha)) \rightarrow [d] \alpha$$

Intuitively we view an accumulator as a “write-only” view of an array. Semantically, accumulators are lists of index/-value pairs, each denoting an update of an array. When we use \( \text{upd} \) on an accumulator, we prepend index/value pair to this list, returning a new accumulator. Operationally, \( \text{upd} \) on an accumulator is implemented by immediately updating the underlying array, not by actually maintaining a list of updates in memory (although such an implementation would be semantically valid). The purpose of accumulators is to allow the compiler to continue to reason purely functionally, in particular that all data dependencies are explicit, while allowing efficient code generation based on in-place updates. Our accumulators are similar to generalized reductions [24] or the accumulation effects in Dex [33] and have the same underlying motivation. The main difference is that in Dex, they are an effect, which requires every part of the compiler to be effect-aware.

Free array-typed variables in \( \text{map} \) are thus turned into accumulators while generating return sweep code for the \( \text{map} \), during which we can perform the updates directly. We allow implicit conversion between accumulators and arrays of accumulators, as this allows us to directly \( \text{map} \) them. E.g.

$$\text{let } \overline{xs} = \text{map} (\lambda i \rightarrow \overline{as}[i]) \text{ is results in the return sweep code} \text{let } \overline{\gamma} = \text{withacc } \overline{\alpha} (\lambda \overline{\alpha}_{\text{acc}} \rightarrow \text{map} (\lambda (i, \overline{x}, \overline{as}) \rightarrow \text{upd} i \overline{x} \overline{as}) \text{ is } \overline{x} \overline{as} \overline{\alpha}_{\text{acc}})$$

where we treat \( \overline{\alpha}_{\text{acc}} \) as an array of accumulators when passed to \( \text{map} \), and treat the result of the \( \text{map} \) as a single accumulator. This is efficient because accumulators have no run-time representation, and saves us from tedious boilerplate.

During the lifetime of the accumulator, the underlying array may not be used—this prevents observation of intermediate state. These rules can be encoded in a linear type system and mechanically checked, which we do in our implementation, but exclude from the paper for simplicity.

Accumulators are sufficient to express the adjoint computation inside maps because (1) any read from an array \( \overline{a}[i] \) is turned into an accumulation on \( \overline{\alpha}[i] \) as discussed at the end of Section 4.2, and (2) the only place on the return sweep where \( \overline{a} \) can be read outside an accumulation statement is the definition of \( \overline{a} \), which by definition is the last use of \( \overline{a} \), hence it is safe to turn it back into a normal variable there.

6 Implementation and Optimizations

We have implemented the reverse AD transformation as a pass in a copy of the publicly available Futhark compiler. The presented transformation rules were tuned to preserve fusion opportunities, both with constructs from the statement’s sweep, and across statements. Section 6.2 discusses several omitted issues, namely how to optimize checkpointing for the arrays that are constructed by in-place updates inside loop nests, and how to support while loops, and Section 6.3 presents the limitations of our implementation.

Since accumulators were not supported in the original language, we have implemented them throughout the compiler—for the GPU backends, they ultimately boil down to atomic updates, such as \( \text{atomicAdd} \) in CUDA. Accumulators however, often result in suboptimal performance, because they access memory in uncoalesced fashion and are subject to conflicts, i.e., threads simultaneously accessing the same location. In this sense, Section 6.1 presents optimizations aimed at turning accumulators into more specialized constructs (e.g., map-reduce) that can be better optimized.

6.1 Optimizing Accumulators

We demonstrate our optimizations of accumulator on the matrix-matrix multiplication. The code below assumes \( \overline{as} \in \mathbb{R}^{m \times q} \) and \( \overline{bs} \in \mathbb{R}^{q \times n} \) and computes \( \overline{cs} \in \mathbb{R}^{m \times n} \) by taking the dot product of each (pair of) row of \( \overline{as} \) and column of \( \overline{bs} \):

$$\text{let } \overline{xs} = \text{map} (\lambda i \rightarrow \text{map} (\lambda j \rightarrow \text{sum} (\text{map} (*) \overline{as}[i, :] \overline{bs}[:, j])) \rightarrow [0..n-1]) \rightarrow [0..m-1]$$

Differentiating the code above results in the return sweep:

$$\text{let } (\overline{as}, \overline{bs}) = \text{withacc } (\overline{as}, \overline{bs}) (\lambda \overline{as}, \overline{bs} \rightarrow \text{map} (\lambda i \overline{bs} \rightarrow \text{map} (\lambda j \overline{bs} \rightarrow \text{map} (\lambda k \overline{bs} \rightarrow \overline{cs}[i, k] = \overline{bs}[k, j] \overline{cs}[i, j] \rightarrow \overline{cs}[i, k] = \overline{as}[i, k] \overline{cs}[i, j]) \overline{bs}[k, j] \overline{cs}[i, j]) \overline{as}[i, k] \overline{cs}[i, j]) \overline{bs}[k, j] \overline{cs}[i, j])$$

12For simplicity we treat only single-dimensional arrays in this section, but the idea also works in the multi-dimensional case. This type for \( \text{withacc} \) allows only a single result corresponding to the array being updated. In practice, we also need to be able to return an arbitrary secondary result.

13https://github.com/diku-dk/futhark
which is not efficient, because (temporal) locality is sub-utilized. In this sense, we have designed and implemented a pass aimed at turning (common cases of) accumulator accesses into reductions. We present the analysis at an intuitive level: The analysis searches for the first accumulator directly nested in a perfect map-nest and checks whether its indices are invariant across any of the parallel dimensions. In our case as is accumulated on indices [i, k] that are both invariant to the outermost parallel index j. In such a case, the map-nest is split into two: the code on which the accumulated statement depends, and the code without the accumulator statement, which is simplified and treated recursively. The map-nest encapsulating the accumulation is reorganized such that the invariant dimension (j) is moved innermost. The accumulated statement is taken out of this innermost map, which is modified to produce (only) the accumulated values, whose sum $\text{reduce} (+\ 0)$ is re-written to be the value accumulated by $\overline{a_b}$:

\[
\text{let } (\overline{a_b}) = \text{withacc}(\overline{a_b} \overline{a_b}) \rightarrow \\
\text{map } (\lambda i \overline{a_b} \rightarrow \text{map } (\lambda k \overline{a_b} \rightarrow \\
\text{let } s = \text{sum } (\text{map } (\times) \text{bs}[i,j] \overline{as}[i,:]) \\
\text{in } \overline{upd} (i,k) s \overline{a_b} \rightarrow (\overline{as}[i,k]) \overline{bs}[k,j] \overline{as}[i,j] \\
\text{)} \text{[0...q-1]} \overline{bs} \text{[0...n-1]} \overline{a_b} \\
\text{let } (\overline{bs}) = \text{withacc}(\overline{bs} \overline{bs}) \rightarrow \\
\text{map } (\lambda k \overline{bs} \rightarrow \text{map } (\lambda j \overline{bs} \rightarrow \\
\text{let } s = \text{sum } (\text{map } (\times) \text{ass}[i,j] \overline{bs}[,:]) \\
\text{in } \overline{upd} (k,j) s \overline{bs} \rightarrow (\overline{bs}[i,k]) \overline{bs}[i,j] \overline{bs}[i,j] \\
\text{)} \text{[0...n-1]} \overline{bs} \text{[0...q-1]} \overline{bs}
\]

One can notice that the code now essentially consists (as expected) of two matrix-multiplication-like kernels. These are optimized by a later pass that performs block and register tiling whenever it finds an innermost map-reduce whose input arrays are invariant to one of the outer-parallel dimensions. We have extended this pass (1) to support accumulators, (2) to keep track of the array layout—i.e., transposed or not, (3) to copy from global to shared memory in coalesced fashion for any layout, and (4) to exploit some of the parallelism of the innermost dot product as inspired from [37].

This optimization is responsible for nearly a one-order-of-magnitude speedup at application level for benchmarks dominated by matrix multiplication such as GMM and LSTM, evaluated in sections 7.6 and 7.7.

6.2 Loop optimizations

As discussed in section 4, by default, loop-variant variables are saved/checkpointed at the entry of each iteration. This technique does not preserve the work asymptotic of the original program when a loop variant array is modified in place. For example, the contrived dependent loop below constructs an array of length $n$ in $O(n)$ work, but the checkpointing of the forward sweep would require $O(n^2)$ work:

\[
\text{loop } (xs) = (xs_0) \text{ for } i = 1...n-1 \text{ do } \\
\text{ xs with } [i,j] = \text{ass}[i,j] + xs[i-1,j]
\]

One can observe however that iteration-level checkpointing is not needed if the loop nest does not exhibit any false dependencies (WAR+WAW): Since no value is “lost” through the loop nest, it is enough to checkpoint $xs$ only once at the entry to the outermost loop of the nest. Moreover, re-execution is safe because all the over-writes are idempotent. In this sense, we allow the user to annotate such loop parameters that are free of false dependencies: we checkpoint them at the loop-nest entry and restore this copy just before entering the return sweep of that nest. Essentially, we consider that false dependencies are properly named as such, and the user should not rely on them, especially in a data-parallel context. Work in the automatic parallelization area can be used to automatically check the safety of such annotations, statically [16], dynamically [11], and anywhere in between [31].

A second issue relates to while loops, on which we cannot perform AD directly, because their statically-unknown iteration count hinders the allocation of the checkpointing tape. We provide two mechanisms to address this issue: first, the user may annotate a while-loop with an iteration bound $n$. The while loop is then transformed into an $n$-iteration for-loop that contains a perfectly nested if, which only executes the valid iterations of the while loop. Alternatively, in the absence of such annotation, the loop count can be computed dynamically by an inspector—i.e., loop slice that computes only the number of iterations of the loop.

6.3 Current Implementation Limitations

We have already discussed limitations during the presentation of individual constructs. The main remaining one is that we do not currently support loop-variant parameters that change their shape throughout the loop. In principle this can be handled by dynamic re-allocations, albeit this might prove expensive in a GPU setting.

Another shortcoming is that when an in-place update occurs inside an if-then-else, we currently save the target array at the branch’s entry: in principle, we should instead propagate the saved element(s)—denoted $xx_{saved}$ in Section 5.3—outside branches, and back in to reach the return sweep. By construction, this propagation cannot escape the scope of the innermost-enclosing recurrence.
7 Experimental Evaluation

7.1 AD-Bench: Sequential AD Performance

ADBench is a collection of benchmarks aimed at comparing different AD tools [43]. We have used the ADBench framework to measure Futhark versions of the four ADBench problems. We compile to sequential CPU code and report the time taken to compute the full Jacobian relative to the time taken to compute the objective function, using the largest default dataset for each benchmark. We compare against Tapenade [2] and manually differentiated programs. The results are shown in Table 1.

Futhark does well, in particular managing to outperform Tapenade in four out of five cases. For the exception, BA, the bottleneck is packing the result (which is a sparse Jacobian) in the CSR format expected by the tooling, which is code that is not subject to AD. The HAND benchmark has two variants: a “simple” one that computes only the dense part of the Jacobian, and a “complicated” one that also computes a sparse part. Tapenade handles the latter poorly. Both Tapenade and Futhark perform poorly when compared to the hand-written code. Both BA and HAND produce sparse Jacobians where the sparsity structure is known in advance, which is exploited by passing appropriate seed vectors to jvp/vjp.

7.2 Parallel Hardware and Methodology

We benchmark on two different Linux systems: one with two AMD EPYC 7352 CPUs and an NVIDIA A100 GPU running CUDA 11.2, which we denote A100, and another with two Intel E5-2650 CPUs and an NVIDIA RTX 2080Ti GPU running CUDA 11.3, which we denote 2080Ti. We report mean runtime for 10 runs, which includes all overheads, except transferring program input and result arrays between device and host. We report the absolute runtime of the differentiated and primal program, and the “overhead” of differentiation that corresponds to the ratio between the two. In optimal AD, this ratio (counted in number of operations) is supposed to be a small constant [15], hence the ratio serves as a good measure of the efficiency of an AD implementation. Futhark benchmarks using 32-bit floats were performed with block and register tile sizes of 16 and 4, respectively; on 64-bit floats the register tile size is 2.

7.3 Compared to Enzyme

RSBench and XSBench are CUDA implementations of Monte Carlo neutron transport algorithms that are reported in the Enzyme paper [29], and which we have ported to Futhark in order to quantitatively compare our solution with Enzyme. They each constitute a large map that contains inner loops and control flow, as well as indirect indexing of arrays. We benchmark on the 2080Ti system, which features a GPU similar to the one used to report the Enzyme results. We use the “small” datasets. The results (Table 2) show the overhead caused by our AD and the one reported in the Enzyme paper. Our overhead is slightly smaller, although this may come down to micro-optimisations.

7.4 Case Study 1: Dense K-means clustering

K-means clustering is an optimization problem where given \( n \) points \( P \) in a \( d \)-dimensional space we must find \( k \) points \( C \) that minimize the cost function

\[
 f(C) = \sum_{p \in P} \min_{c \in C} (\|p - c\|, c \in C)
\]

Solving this with Newton’s Method requires computing the Jacobian and Hessian of \( f \). The cost function is easily written with nested map and reduce operations, for which the Hessian can be computed by nesting vjp inside of jvp. Note that while

\[
 f : \mathbb{R}^{k \times d} \to \mathbb{R}
\]

and so needs reverse-mode vjp, the differentiated function

\[
 vjp \ f : \mathbb{R}^{k \times d} \to \mathbb{R} \to \mathbb{R}^{k \times d}
\]

is suitable for use with forward-mode jvp (the \( \mathbb{R} \) argument is set to 1). Further, as the Hessian for \( f \) has nonzero entries

<table>
<thead>
<tr>
<th>Tool</th>
<th>BA</th>
<th>D-LSTM</th>
<th>GMM</th>
<th>HAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Futhark</td>
<td>13.0×</td>
<td>3.2×</td>
<td>5.1×</td>
<td>49.8×</td>
</tr>
<tr>
<td>Tapenade</td>
<td>10.3×</td>
<td>4.5×</td>
<td>5.4×</td>
<td>3758.7×</td>
</tr>
<tr>
<td>Manual</td>
<td>8.6×</td>
<td>6.2×</td>
<td>4.6×</td>
<td>4.6×</td>
</tr>
</tbody>
</table>

Table 1. Time to compute the full Jacobian relative to time to compute the objective function. Lower is better.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Primal runtimes</th>
<th>AD overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSBench</td>
<td>2.311s</td>
<td>3.6×</td>
</tr>
<tr>
<td>XSBench</td>
<td>0.244s</td>
<td>2.6×</td>
</tr>
</tbody>
</table>

Table 2. Results for RSBench and XSBench. We report the absolute runtimes of the original program and the un-differentiated Futhark program, as well as the overhead of a single forward and return sweep of the reverse-mode differentiated program compared to the un-differentiated one.

<table>
<thead>
<tr>
<th>(k, n, d)</th>
<th>Futhark</th>
<th>AD</th>
<th>PyTorch</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5, 494019, 35)</td>
<td>(1024, 10000, 256)</td>
<td>17.8ms</td>
<td>72.6ms</td>
</tr>
<tr>
<td>(18.8ms)</td>
<td>22.6ms</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Performance measurements for k-means clustering for two different workloads, implemented both manually in Futhark, with AD in Futhark, and with AD in PyTorch.
only along the diagonal, we need only a single invocation of \( \text{jvp} \) to compute it, returning exactly this diagonal. This shows how \( \text{jvp/vjp} \) allows the user to exploit sparsity. Further, the nesting of \( \text{jvp/vjp} \) allows us to avoid duplicating the expensive search operation in the cost function, which would not be the case if we were limited to dedicated constructs for Jacobian and Hessian computation.

\( k \)-means clustering can also be solved using a method based on repeatedly calculating histograms [17], which can be implemented using Futhark’s dedicated histogram construct. The code that is generated by AD consists of maps that perform semantically equivalent histogram-like accumulator updates. Essentially, using AD leads to basically the same code one would write by hand. The main difference is that in the hand-written code, some of the updates have are manually sequentialized, while the AD version maximizes parallelism to degree that is detrimental for some datasets.

In PyTorch we must take care when computing the all-pairs Euclidean distance from data points to the centers. Expressing this computation with broadcasting suffers from massive memory overhead. We get around this by explicitly distributing the quadratic terms in the Euclidean distance. The AutoGrad module computes the Jacobian and Hessian, instead of in one go as we can when nesting \( \text{jvp/vjp} \).

We benchmark with two qualitatively different datasets, with results in Table 3. When the histograms benefit from the optimisations discussed in [17], the hand-written implementation has a speedup of 4\( \times \) over our AD approach, while there is no significant difference if they cannot. Our AD approach is slightly faster than PyTorch (as high as 1.22\( \times \) on \textbf{A100} and 1.18\( \times \) on \textbf{2080Ti}).

### 7.5 Case Study 2: Sparse \( k \)-means clustering

We have implemented and tested a sparse formulation of \( k \)-means, that uses a dense representation for centroids and a sparse representation of data. Futhark implementations are relatively straightforward adaptations of the dense \( k \)-means to use the CSR format instead.

The PyTorch implementation uses the COO format because AD (\( \text{jvp} \) and \( \text{hvp} \)) currently raises runtime errors with CSR format. To efficiently compute the cost function we expand the all pairs norm between data \( d \) and centroids \( c \):

\[
||d_i - c_j||^2 = d_i^2 + c_j^2 - 2 \cdot d_i c_j^T. 
\]

In expanded form, all terms can be computed using vectorized operations. Due to the array (tensor) presentation \( d_i^2 \) and \( d_i c_j^T \) are computed using PyTorch specialized sparse module, in particular, the (sparse) matrix product uses \text{sparse.mm} which allows sparse gradient computation (but also forces the COO representation).

Table 4 shows the runtime results for \textbf{A100} on three publicly available NLP workloads for \( k = 10 \). Our AD is slower than the manual code by a factor between 2.5 – 3.7\( \times \) because the generalized-histogram implementation of the manual code uses a multi-pass technique that allows the updates to fit in the L2 cache [17]. PyTorch AD is more than 400\( \times \) slower than our AD, which we do not know how to explain.

### 7.6 Case Study 3: GMM

To evaluate the parallelism-preservation of our AD transformation, we compile the GMM benchmark from the AD-Bench suite to parallel CUDA. We compare against AD-Bench’s implementation of GMM in PyTorch (also run on CUDA), which we have improved (by a > 10\( \times \) factor) by vectorizing all comprehensions. We benchmark on a selection of 1,000 and 10,000-point datasets from AD-Bench featuring a variety of \( d \) (the dimensionality of the input data) and \( K \) (the number of Gaussian distributions used in the model) values.

<table>
<thead>
<tr>
<th>Workload</th>
<th>Futhark</th>
<th>PyTorch</th>
</tr>
</thead>
<tbody>
<tr>
<td>movienews</td>
<td>Manual</td>
<td>61ms</td>
</tr>
<tr>
<td></td>
<td>AD</td>
<td>152ms</td>
</tr>
<tr>
<td></td>
<td>PyTorch</td>
<td>61223ms</td>
</tr>
<tr>
<td>nytimes</td>
<td>Manual</td>
<td>83ms</td>
</tr>
<tr>
<td></td>
<td>AD</td>
<td>300ms</td>
</tr>
<tr>
<td></td>
<td>PyTorch</td>
<td>226896ms</td>
</tr>
<tr>
<td>scrna</td>
<td>Manual</td>
<td>156ms</td>
</tr>
<tr>
<td></td>
<td>AD</td>
<td>579ms</td>
</tr>
<tr>
<td></td>
<td>PyTorch</td>
<td>367799ms</td>
</tr>
</tbody>
</table>

(\textbf{a}) Parameters of the AD-Bench datasets used for benchmarking. The corresponding datasets may be found on the AD-Bench GitHub repository (https://github.com/microsoft/ADBench).

<table>
<thead>
<tr>
<th></th>
<th>PyTorch Jacob. (ms)</th>
<th>PyT. Jacob. (ms)</th>
<th>Fut. Speedup (( \times ))</th>
<th>Fut. Overhead (( \times ))</th>
<th>PyT. Overhead (( \times ))</th>
<th>Fut. Overhead (( \times ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.6</td>
<td>19.2</td>
<td>12.2</td>
<td>1.87</td>
<td>1.81</td>
<td>1.81</td>
</tr>
<tr>
<td></td>
<td>1.85</td>
<td>2.18</td>
<td>1.14</td>
<td>1.71</td>
<td>1.69</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>2.64</td>
<td>5.28</td>
<td>2.09</td>
<td>3.09</td>
<td>3.04</td>
<td>3.04</td>
</tr>
<tr>
<td></td>
<td>2.34</td>
<td>2.20</td>
<td>2.21</td>
<td>2.12</td>
<td>2.23</td>
<td>2.23</td>
</tr>
</tbody>
</table>

(\textbf{b}) GMM benchmark results on the \textbf{A100} system with 64-bit floats. \textit{Fut.} and \textit{PyT.} refer to Futhark and PyTorch. \textbf{PyT. Jacob.} is the time to compute the full Jacobian of the objective function in PyTorch.

<table>
<thead>
<tr>
<th></th>
<th>PyT. Jacob. (ms)</th>
<th>PyT. Jacob. (ms)</th>
<th>Fut. Speedup (( \times ))</th>
<th>Fut. Overhead (( \times ))</th>
<th>PyT. Overhead (( \times ))</th>
<th>Fut. Overhead (( \times ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20.2</td>
<td>57.9</td>
<td>28.7</td>
<td>31.1</td>
<td>17.5</td>
<td>17.5</td>
</tr>
<tr>
<td></td>
<td>4.5</td>
<td>5.29</td>
<td>1.45</td>
<td>7.31</td>
<td>6.20</td>
<td>6.20</td>
</tr>
<tr>
<td></td>
<td>2.28</td>
<td>2.41</td>
<td>1.07</td>
<td>3.20</td>
<td>2.62</td>
<td>2.62</td>
</tr>
<tr>
<td></td>
<td>2.8</td>
<td>2.5</td>
<td>2.48</td>
<td>2.36</td>
<td>3.27</td>
<td>3.27</td>
</tr>
</tbody>
</table>

(\textbf{c}) GMM benchmark results on the \textbf{2080Ti} system with 32-bit floats. The PyTorch implementation ran out of memory on the \textbf{D3} dataset, indicated with \( \emptyset \). The Futhark Jacobian-runtime was 96.7\( ms \) on \textbf{D5}.

<table>
<thead>
<tr>
<th></th>
<th>PyT. Jacob. (ms)</th>
<th>PyT. Jacob. (ms)</th>
<th>Fut. Speedup (( \times ))</th>
<th>Fut. Overhead (( \times ))</th>
<th>PyT. Overhead (( \times ))</th>
<th>Fut. Overhead (( \times ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.6</td>
<td>19.2</td>
<td>12.2</td>
<td>1.87</td>
<td>1.81</td>
<td>1.81</td>
</tr>
<tr>
<td></td>
<td>1.85</td>
<td>2.18</td>
<td>1.14</td>
<td>1.71</td>
<td>1.69</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>2.64</td>
<td>5.28</td>
<td>2.09</td>
<td>3.09</td>
<td>3.04</td>
<td>3.04</td>
</tr>
<tr>
<td></td>
<td>2.34</td>
<td>2.20</td>
<td>2.21</td>
<td>2.12</td>
<td>2.23</td>
<td>2.23</td>
</tr>
</tbody>
</table>

Table 5. Benchmarking results for the GMM case study.
see Table 5a. We observed that the runtime of the primal program is dominated by matrix multiplication (~70%).

Matrix multiplication is a primitive in PyTorch [32]; we reasonably expect that differentiation of matrix multiplication may be readily implemented very efficiently. In Futhark there are no such primitives: matrix multiplication is written with maps, whose differentiation yield accumulators, which are further optimized as described in Section 6.1.

The benchmark results are shown in Tables 5b and 5c for the A100 and 2080Ti systems, respectively. On the A100 64-bit floats were used; on the 2080Ti system 32-bit floats were used due to the limited double-precision performance. The results on 2080Ti demonstrate a significant speedup of Futhark over PyTorch (as high as 7.3×) in the Jacobian runtimes on all inputs. Futhark also gets the upper hand on A100, albeit the average speedup is smaller (1.68×).

<table>
<thead>
<tr>
<th></th>
<th>PyT. Jacob.</th>
<th>Speedups (×)</th>
<th>Overheads (×)</th>
<th>PyT. Fut. cuDNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_0$</td>
<td>51.9ms</td>
<td>3.1 14.0</td>
<td>2.6 2.0 2.8</td>
<td>3.6 3.0 3.2</td>
</tr>
<tr>
<td>$D_1$</td>
<td>713.7ms</td>
<td>3.0 25.5</td>
<td>3.6 4.0 1.7</td>
<td>2.6 2.6 3.2</td>
</tr>
<tr>
<td>$D_0$</td>
<td>76.8ms</td>
<td>3.1 8.3</td>
<td>2.6 2.6 3.2</td>
<td>3.2 3.2 3.2</td>
</tr>
<tr>
<td>$D_1$</td>
<td>1175.0ms</td>
<td>3.2 12.8</td>
<td>2.5 4.0 2.5</td>
<td>3.2 3.2 3.2</td>
</tr>
</tbody>
</table>

Table 6. LSTM measurements on datasets $D_0$: ($bs, n, d, h$) = (1024, 20, 300, 192) and $D_1$: ($bs, n, d, h$) = (1024, 300, 80, 256). PyT., Fut., and cuDNN refer to the PyTorch, Futhark, and the cuDNN-based torch.nn.LSTM implementations.

7.7 Case Study 3: LSTM

Long Short-Term Memory (LSTM) [40] is a type of recurrent neural network architecture popular in named entity recognition and part-of-speech tagging [10] [38]. We benchmark two LSTM networks with hyperparameters common in natural language processing. The LSTM’s input dimensions ($d$) are 80 and 300, respectively. These correspond to GloVe embeddings [36]. The recurrent units have hidden dimensions ($h$) 192 and 256 [10]. The sequence lengths ($n$) are 20 and 300. The lengths are close to the average sentence length for the Penn Treebank dataset [27] and IMDB dataset [25]. Batch sizes ($bs$) were chosen to fully saturate the GPUs.

Both the Futhark and PyTorch implementations are based on the architecture in [40]. We also compare against PyTorch’s torch.nn.LSTM class, which wraps the NVIDIA cuDNN LSTM implementation [9]; note that this implementation is hand-written/optimized and features manual differentiation. All benchmarks were on 32-bit floating points and were initialized with the same parameter data, which was generated by torch.nn.LSTM. The results are shown in Table 6.

Futhark is about 3× faster than PyTorch on both 2080Ti and A100 systems. As with GMM, LSTM is dominated by matrix-multiplication computations. Neither AD-based implementation is competitive against the manual cuDNN-based implementation—we suspect that torch.nn.LSTM switches the computation to tensor cores (F16) internally, because the speedups over Futhark are well correlated with the ratio between peak F32 and F16 performance of the two systems.

8 Related Work

Reverse-mode differentiation of reduce and scan is discussed in [34]. Our rule for reduce is similar, but we handle scan differently. Our approach is less efficient for complex operands because we manifest the Jacobian matrices, but more efficient for single-value operands on GPUs as it requires less shared memory to implement the derived scan operator. Neither our scan rule nor the one from [34] is asymptotics-preserving in the general case, but they are for most scans that occur in practice.

$F$ is a functional array language that supports nested parallelism. Its AD implementation uses the forward mode, along with a handful of rewrite rules that allow it to exploit sparsity in some cases [41].

Dex is a recent language built specifically to support efficient AD. Empirical benchmarks for AD in Dex have not yet been published, but we can compare their approach [33]. In contrast to our conventional “monolithic” approach where reverse-mode AD is a transformation completely distinct from forward-mode, Dex uses a technique where the program is first “linearized”, producing a linear map, after which this linear map is then “transposed”, producing the adjoint code. Like Dex, we do not support recursion or AD of higher-order functions. Dex does not make direct use of a “tape” in the classical sense, but instead constructs arrays of closures followed by defunctionalization. The actual run-time data structures will conceptually consist of multiple tapes in the form of multidimensional irregular arrays. Dex does not yet make use of checkpointing, or optimization of particular accumulation patterns as in Section 6.1.

Enzyme shows the advantage of performing AD after standard compiler optimisations has simplified the program [28]. Like Enzyme, we also apply our AD transformation on a program that has already been heavily optimized by the compiler. But where Enzyme is motivated by performing AD on a post-optimization low-level representation, our work takes advantage of both pre-AD optimization, as well as the information provided by high-level parallel constructs. Enzyme has also been applied to GPU kernels [29]. We achieve equivalent performance, but our approach is not based on differentiating single kernels—indeed, the GPU code we generate for a differentiated program may have a significantly different structure than the original program. For example, the optimized adjoint code for a matrix multiplication requires two matrix multiplications, each its own kernel, as in the LSTM and GMM benchmarks.
Recent AD work on OpenMP details an approach to reverse-mode AD for non-nested parallel loops [22]. As with map, the chief difficulty is handling adjoint updates to free variables in the loop. Two approaches are used: a) the variable is copied across threads and per-thread updates to the adjoint will be combined at loop termination or b) the variable is shared across threads and updates are handled via atomic operations. Copying free variables across threads does not in general preserve the asymptotics of the original program when the free variables are arrays; shared variables avoid this issue, but instead can result in significant sequentialization. Due to our redundant recomputation technique (avoiding the difficulty of passing a tape across scopes), our implementation supports and preserves nested parallelism in the original program. Hence, while our implementation is also realized by atomic operations, we are able to exploit nested parallelism to identify and reduce some updates into a single atomic update and thereby leverage benefits from both approaches.

ML practitioners use tools such as PyTorch [32] that incorporate AD. These are less expressive than our language and do not support true nested parallelism, but instead require the program to use flat (although vectorized) constructs. On the other hand, they can provide hand-tuned adjoints for the primitives they do support. JAX is another such example: it supports automatic differentiation of pure Python code and just-in-time (JIT) compilation to XLA HLO [8, 13]. However, JAX does not support reverse-mode differentiation of their loop primitive (for i in range) when the iteration count of the loop isn’t statically known. JAX’s JIT approach requires re-interpretation (by Python) and compilation (by XLA) when functions are called on arguments with so far unseen shape/underlying type, but this approach may enable better code specialization than our entirely ahead-of-time approach.

Reverse AD has also been implemented in DSLs aimed at stencil computations [23]. The challenge here is to combine AD with loop transformations such that the resulted code is a stencil itself, and thus it can be optimized with the repertoire of existent optimizations. AD has also been described for tensor languages that support only constrained forms of loops, which in particular has the benefit of not requiring the use of tapes [4].

9 Conclusions

We have presented a fully operational compiler implementation of both reverse and forward mode AD in a nested-parallel functional language. Our experimental evaluation shows that our mapping of parallel construct and our technique of implementing the tape based on redundant computation is practically effective, and competitive with both (1) well-established frameworks that encompass more specialized languages such as PyTorch, and with (2) newer research efforts aimed at a lower-level language, such as Enzyme.

References


