

Instruction level automatic differentiation on a reversible Turing machine

Jin-Guo Liu,^{1,*} Taine Zhao,² and [JG: MLS's name here]³

¹*Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China*

²*Department of Computer Science, University of Tsukuba*

³*MLS's affiliations here*

This paper considers source to source automatic differentiation (AD) on a reversible Turing machine. We start by reviewing why adjoint mode AD is hard for traditional machine learning frameworks and propose a solution to existing issues by writing a program reversibly. We developed an reversible eDSL NiLang in Julia that can be used to generate backward rules. We demonstrate its power by differentiating over the reversible implementations of Bessel function, and linear algebra functions unitary matrix multiplication and QR decomposition. It is also a promising direction towards solving the notorious memory wall problem in machine learning. We also discuss the challenges that we face towards rigorous reversible programming from the instruction and hardware perspective.

I. INTRODUCTION

Computing the gradients of a numeric model $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ plays a crucial role in scientific computing. Consider a computing process

$$\begin{aligned} \mathbf{x}^1 &= f_1(\mathbf{x}^0) \\ \mathbf{x}^2 &= f_2(\mathbf{x}^1) \\ &\dots \\ \mathbf{x}^L &= f_L(\mathbf{x}^{L-1}) \end{aligned}$$

where $\mathbf{x}^0 \in \mathbb{R}^m$, $\mathbf{x}^L \in \mathbb{R}^n$, L is the depth of computing. The Jacobian of this program is a $n \times m$ matrix $J_{ij} \equiv \frac{\partial x_j^L}{\partial x_i^0}$, where x_j^0 and x_i^L are single elements of inputs and outputs. Computing part of the Jacobian automatically is what we called automatic differentiation (AD). It can be classified into three classes, the tangent mode AD, the adjoint mode AD and the mixed mode AD. [1] The tangent mode AD computes the Jacobian matrix elements that related to a single input using the chain rule $\frac{\partial \mathbf{x}^k}{\partial x_j^0} = \frac{\partial \mathbf{x}^k}{\partial \mathbf{x}^{k-1}} \frac{\partial \mathbf{x}^{k-1}}{\partial x_j^0}$, while a tangent mode AD computes Jacobian matrix elements that related to a single output using $\frac{\partial \mathbf{x}^k}{\partial x_j^0} = \frac{\partial \mathbf{x}^k}{\partial \mathbf{x}^{k-1}} \frac{\partial \mathbf{x}^{k-1}}{\partial x_j^0}$. In variational applications where the loss function always outputs a scalar, the adjoint mode AD is preferred. However, implementing adjoint mode AD is harder than implementing its tangent mode counterpart, because it requires propagating the gradients in the inverse direction of computing the loss. The back propagation of gradients requires intermediate information of a program that includes

1. the computational process,
2. and variables used in computing gradients.

The computational process is often stored in a computational graph, a directed acyclic graph (DAG) that represents the relationship between data and functions. In Pytorch [2] and Flux [3], every variable has a tracker field that stores its

parent information, i.e., the input data and function generating this variable. TensorFlow [4] implements a static computational graph as a description of the program before actual computation happens. The required variables are also recorded in this graph. For source to source AD package, Tools like Tapenade [1] and Zygote [5, 6] generate the adjoint code statically while putting variable on a stack called the wengert list. [JG: The introduction to source to source AD is a bit weak, please give reader an overview about the development and difficulty in this field.] As reverse mode AD require storing intermediate result, the memory consumption of the program is typically $O(n)$ where n is the runtime of the program. In many deep learning models like recurrent neural network [7] and residual neural networks [8], the depth can reach several thousand, where the memory is often the bottleneck of these programs. In traditional AD application, program that might do billions of computations will as well get a wengert list in the range of GBs, thus memory will become a bottleneck as well. The memory management is becoming a wall [9] that limits the scale of many applications on GPU. The computational power of a Nvidia V100 GPU can reaches 100 TFLOPS, which is the scale of a mini cluster. However, it memory is only 32GB. A better solution to memory management should be found. [JG: @MLS, Do you have any good comments about starving CPU, i.e. memory accessing is consuming much more time than computing itself. This is why caching intermediate result is a performance killer.]

We tackle this problem by making a program reversible. In the machine learning field, reversibility has been used in reduce the memory allocations in recurrent neural networks [10] and residual neural networks [11]. These techniques include information buffer [12] and reversible activation functions [13, 14]. However, our approach is general purposed. We develop an embedded domain-specific language (eDSL) NiLang in Julia language [15, 16] that implements reversible programming. [17, 18]. This eDSL provides a macro to generate reversible functions, and is completely compatible with Julia ecosystem. One can write reversible control flows, instructions and memory managements in this macro. Combining it with Julia's type system, we implement the AD engine within 100 lines that differentiate any program

* cagate0129@iphy.ac.cn

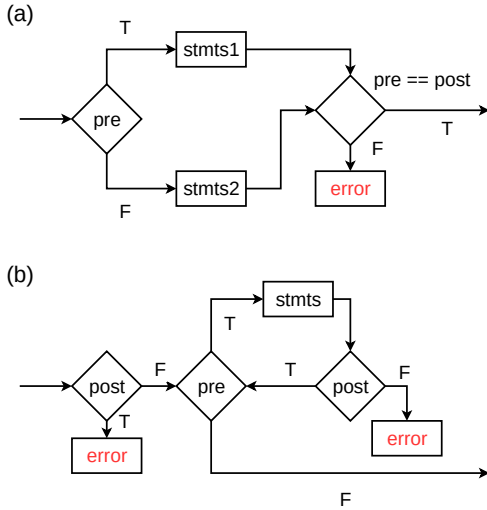


Figure 2: Flow chart for reversible (a) if statement and (b) while statement. “stmts”, “stmts1” and “stmts2” are statements, statements in true branch and statements in false branch respectively. “pre” and “post” are precondition and postconditions respectively.

i.e. additional space. This is a trade off between space and time, which we will discuss the detail in Sec. V A.

2. Control flows

The reversible if statement is shown in Fig. 2 (a). It contains two conditions, a precondition and a postcondition. The precondition decides which branch to enter in the forward execution, while the postcondition decides which branch to enter in the backward execution. After executing the specific branch, the program checks the consistency between precondition and postcondition to make sure they are the consistent. The reversible while statement is shown in Fig. 2 (b). It also has both precondition and postcondition. Before executing the condition expressions, the program preassumes the postcondition is false. After each iteration, the program asserts the postcondition to be true. In the reverse pass, we exchange the precondition and postcondition. The reversible for statement is similar to irreversible ones except after executing the loop, the program checks the values of these variables to make sure they are not changed. In the reverse pass, we exchange start and stop and inverse the sign of step.

3. Arithmetic instructions

Every instruction has a unique inverse that can undo the changes. For logical operations, we have $y \vee= f(\text{args} \dots)$ self reversible. For other arithmetic operations, we assume $y += f(\text{args} \dots)$ and $y -= f(\text{args} \dots)$ are reversible to each other although it is not true for floating point numbers due to the rounding error. Here f can be identity, $*$, $/$ and

\wedge et. al. Besides the above two types of operations, SWAP operation that exchanges the contents in two memory spaces are also widely used in reversible computing systems. We will have a detailed discussion about the rounding error and number system later in Sec. V B

B. NiLang’s Reversible IR

In the last subsection, we have reviewed basic building blocks of a typical reversible language. In order to insert the code of obtaining gradients into the reversed program, the reversible language design should have related abstraction power. This motivates us to design a new reversible language NiLang to fit this task. NiLang is an eDSL in Julia. We choose Julia as the host language for multiple purposes. Julia’s meta programming and its package for pattern matching MLStyle.jl [34] allow us to define an eDSL conveniently. Meanwhile, the type inference and just in time compiling can remove most overheads introduced in our eDSL, providing a reasonable performance. Most importantly, the multiple dispatch provides the polymorphism that will be used in our autodiff engine.

The main feature of NiLang is contained in a single macro @i that compiles a reversible function. The allowed statements in this eDSL are shown in Appendix A. The following is a minimal example of compiling a NiLang function to native julia function.

```
julia> using NiLangCore, MacroTools

julia> macroexpand(Main, :(@i function f(x, y)
    SWAP(x, y)
end)) |> MacroTools.prettify
quote
  $(Expr(:meta, :doc))
  function $(Expr(:where, :(f(x, y))))
    gaur = SWAP(x, y)
    x = (NiLangCore.wrap_tuple(gaur))[1]
    y = (NiLangCore.wrap_tuple(gaur))[2]
    return (x, y)
  end
  if typeof(f) != typeof(~f)

    function $(Expr(:where, :((# $ TODO: remove this comment
      mongoose::typeof(~f))(x, y))))
      mandrill = (~SWAP)(x, y)
      x = (NiLangCore.wrap_tuple(mandrill))[1]
      y = (NiLangCore.wrap_tuple(mandrill))[2]
      return (x, y)
    end
  end
  if !(NiLangCore._hasmethod1(
    NiLangCore.isreversible, typeof(f)))
    NiLangCore.isreversible(::typeof(f)) = true
  end
end
```

Macro @i generates three functions f , $\sim f$ and $\text{NiLangCore.isreversible}$. f and $\sim f$ are a pair of

functions that reverse to each other, where $\sim f$ is an callable of type `Inv{typeof(f)}`. In the body of `f`, `NiLangCore.wrap_tuple` is used to unify output data types, it will wrap any non-tuple variable to a tuple. The outputs of `SWAP` are assigned back to its input variables, in other words, a function modifies inputs inplace. At the end this function, this macro attaches a return statement that returns all input variables. `NiLangCore.isreversible` is a function to mark the reversibility trait of `f`.

To understand the design of reversibility, we first introduce a reversible IR that plays a central role in `NiLang`. In this IR, a statement can be an instruction, a function call, a control flow, a memory allocation/deallocation, or the inverse statement “ \sim ”. Any statement in this IR has a unique inverse as shown in Table I.

statement	inverse
<code><f>(<args>...)</code>	<code>(~<f>)(<args>...)</code>
<code><y> += <f>(<args>...)</code>	<code><y> -= <f>(<args>...)</code>
<code><y> .+= <f>(<args>...)</code>	<code><y> .-= <f>(<args>...)</code>
<code><y> ∇= <f>(<args>...)</code>	<code><y> ∇= <f>(<args>...)</code>
<code><y> .∇= <f>(<args>...)</code>	<code><y> .∇= <f>(<args>...)</code>
<code><a> ← <expr></code>	<code><a> → <expr></code>
<code>(<T1> => <T2>)(<x>)</code>	<code>(<T2> => <T1>)(<x>)</code>
<code>begin <stmts> end</code>	<code>begin ~(<stmts>) end</code>
<code>if (<pre>, <post>) <stmts1> else <stmts2> end</code>	<code>if (<post>, <pre>) ~(<stmts1>) else ~(<stmts2>) end</code>
<code>while (<pre>, <post>) <stmts> end</code>	<code>while (<post>, <pre>) ~(<stmts>) end</code>
<code>for <i>=<m>:<s>:<n> <stmts> end</code>	<code>for <i>=<m>:-<s>:<n> ~(<stmts>) end</code>
<code>@safe <expr></code>	<code>@safe <expr></code>

Table I: A collection of reversible statements. “ \cdot ” is the symbol for broadcasting magic in Julia, “ \sim ” is the symbol for reversing a statement or a function. `<...>` represents a non-keyword, where `<pre>` stands for precondition, `<post>` stands for postcondition, `<args>...` stands for the argument list of a function, `<stmts>` stands for statement, `<exprs>` stands for expression, `<T1>` and `<T2>` stand for types and the reset are variables.

“ \leftarrow ” and “ \rightarrow ” are symbols for memory allocation and deallocation, one can input them by typing “`\leftarrow`” and “`\rightarrow`” respectively followed by a Tab key in a Julia editor or REPL. “`begin <stmts> end`” is the block statement in Julia, it represents a code block. It can be inverted by reversing the order of `<stmts>` as well as each element in

it. The conditional expression in `if` or `while` statements is a tuple of precondition and postcondition. Finally, the special macro `@safe` allows users to use external statements that do not break reversibility. For example, one can use `@safe @show <var>` for debugging.

C. Compiling

The compilation of a reversible function contains three stages.

The first stage preprocess human inputs to a reversible IR. The preprocessor expands the symbol “ \sim ” in the postcondition field of `if` statement by copying the precondition, adds missing ancilla “ \leftarrow ” statements to ensure “ \leftarrow ” and “ \rightarrow ” appear in pairs inside a function, a `while` statement or a `for` statement, and expands the uncomputing macro `~@routine`. Since the compute-copy-uncompute paradigm is extensively used in reversible programming for uncomputing ancillas. One can use `@routine <stmt>` statement to record a statement, and `~@routine` to insert `~<stmt>` for uncomputing. The following example preprocesses an `if` statement to the reversible IR.

```
julia> using NiLangCore, MacroTools

julia> MacroTools.prettify(
    @code_preprocess if (x > 3, ~)
        @routine z += x * y
        ~@routine
    end)
:(if (x > 3, x > 3)
    z += x * y
    z -= x * y
else
end)
```

In this example, since the precondition “`x > 3`” is not change after execution of the specific branch, we omit the postcondition by putting a “ \sim ” in this field. “`@routine`” records a statement, the statement can also be a “`begin <stmts> end`” block as a sequence of statements.

The second stage generates the reversed code according to table Table I. For example,

```
julia> MacroTools.prettify(
    @code_reverse if (pre, post)
        z += x * y
    else
        z += x / y
    end)
:(if (post, pre)
    z -= x * y
else
    z -= x / y
end)
```

The third stage is translating this IR and its inverse to native Julia code. It explains all functions as inplace and insert codes about reversibility check. At the end of a function definition, it attaches a return statement that returns all input arguments. After this, the function is ready to execute on the host language. The following example shows how an `if` statement is transformed in this stage.

```
julia> MacroTools.prettify(
    @code_interpret if (pre, post)
        z += x * y
    else
        z += x / y
    end)
quote
    bat = pre
    if bat
        @assignback (PlusEq(*))(z, x, y)
    else
        @assignback (PlusEq(/))(z, x, y)
    end
    @invcheck post bat
end
```

The compiler translates the instruction according to Table II and adds `@assignback` before each instruction and function call statement. The macro `@assignback` assigns the output of a function back to the arguments of that function. `@invcheck post bat` checks the consistency between preconditions and postconditions to ensure reversibility. This statement will throw an `InvertibilityError` error if target variables `bat` and `post` are not “equal” to each other up to a certain tolerance.

D. Types and Dataviews

So far, the language design is not too different from a traditional reversible language. To implement the adjoint mode AD, we introduce types and dataviews. The type that used in the reversible context is just a normal Julia type with an extra requirement of having reversible constructors. The inverse of a constructor is called a “destructor”, which unpacks data and deallocates derived fields. Data packing is

implemented by reinterpreting the new function in Julia. For example,

```
x ← new{TX, TG}(x, g)
```

Here, the “←” statement followed by a new function is treated specially that it deallocates `g`. This makes sense because the output of `new` keeps all information in input argument list. Its inverse is

```
x → new{TX, TG}(x, g)
```

It unpacks `x` and allocates a new ancilla `g`. The following example shows how to define a reversible type `GVar`.

```
julia> using NiLangCore

julia> @i struct GVar{T,GT} <: IWrapper{T}
    x::T
    g::GT
    function GVar{T,GT}(x::T, g::GT) where {T,GT}
        new{T,GT}(x, g)
    end
    function GVar(x::T, g::GT) where {T,GT}
        new{T,GT}(x, g)
    end
    @i function GVar(x::T) where T
        g ← zero{GT}()
        x ← new{T,T}(x, g)
    end
    @i function GVar(x::AbstractArray)
        GVar.(x)
    end
end

julia> GVar{Float64}(0.5)
GVar{Float64,Float64}(0.5, 0.0)

julia> (~GVar)(GVar{Float64}(0.5))
0.5

julia> (~GVar)(GVar{Float64}([0.5, 0.6]))
2-element Array{Float64,1}:
 0.5
 0.6
```

This piece of code is copied from the `autodiff` submodule of `NiLang`. `GVar` is a type used to store gradient information of a variable. Here, we put `@i` macro before both `struct` and `function` statements. The ones before functions mark reversible functions, while the one before `struct` keyword is used to handle the function scope issue. It moves `~GVar` functions outside of this type definition.

The reversible cast between two types can be defined

conveniently with the macro `@icast`.

```
julia> @pure_wrapper A

julia> @icast A(x) => GVar(x, g) begin
    g ← zero(x)
    g += identity(1)
end

julia> x = A(0.5)
A(0.5)

julia> @instr (A=>GVar)(x)
GVar{Float64,Float64}(0.5, 1.0)

julia> @instr (GVar=>A)(x)
A(0.5)
```

Here, we first define a simple reversible wrapper `A` using macro `@pure_wrapper`, and then the cast rule between `A` type and `GVar` type. The body of cast is a reversible mapping that transforms `x` to `(x, g)`. The compiler appends a default constructor `DVar(xx, gg)` at the end of program to instantiate a new object as the return value. Its inverse that converts an object of type `GVar` to type `A` is automatically generated by reversing the above statements.

The fields of an object can be accessed and altered through dataviews. A dataview can be an object, a field of a dataview, an array element of a dataview, or a bijective mapping of a dataview. Let us first consider the following example.

```
julia> arr = [GVar(3.0), GVar(1.0)]
2-element Array{GVar{Float64,Float64},1}:
 GVar{Float64,Float64}(3.0, 0.0)
 GVar{Float64,Float64}(1.0, 0.0)

julia> x, y = 1.0, 2.0
(1.0, 2.0)

julia> @instr -arr[2].g += x * y
2.0

julia> arr
2-element Array{GVar{Float64,Float64},1}:
 GVar{Float64,Float64}(3.0, 0.0)
 GVar{Float64,Float64}(1.0, -2.0)
```

Here, both `-arr[2].g`, `x` and `y` are dataviews. In Julia language, the statement `-grad(arr[2]) += x * y` should throw a syntax error because the function call “-” can not be assigned, and `GVar` is an immutable type. In our eDSL, we wish it works because a memory cell is assumed to be modifiable in our eDSL. The secret of how it works lies in the macro `@assignback`, it translates the above statement to

```
1 res = (PlusEq(*))(-arr[2].g, x, y)
2 arr[2] = chfield(arr[2], Val{:g},
3   chfield(arr[2].g, -, res[1]))
4 x = res[2]
5 y = res[3]
```

The first line `PlusEq(*)(-arr[2].g, x, y)` computes the output, which is a tuple of length 3. At lines 2-3, `chfield(x, Val{:g}, val)` modifies the `g` field of `x` and `chfield(x, -, res[1])` returns `-res[1]`. Here, modifying a field requires the default constructor of a type not overwritten. The assignments in lines 4 and 5 are straightforward.

III. AUTOMATIC DIFFERENTIATION

Local Jacobians and Hessians for basic instructions used in this section could be found in Appendix B 1.

A. First order gradient

Consider a computation $\mathbf{x}^{i-1} = f_i^{-1}(\mathbf{x}^i)$ in a reversed program, the Jacobians can be propagated in the reversed direction like [MLS: adjoint mode AD does not propagate Jacobians.][JG: now?]

$$\begin{aligned} J_{\mathbf{x}^{L'}}^{\mathbf{x}^L} &= \delta_{\mathbf{x}^L, \mathbf{x}^{L'}}, \\ J_{\mathbf{x}^{i-1}}^{\mathbf{x}^L} &= J_{\mathbf{x}^i}^{\mathbf{x}^L} J_{\mathbf{x}^{i-1}}^{\mathbf{x}^i}, \end{aligned} \quad (1)$$

where \mathbf{x}^L represents the outputs of the program, $J_{\mathbf{x}^i}^{\mathbf{x}^L} \equiv \frac{\partial \mathbf{x}^L}{\partial \mathbf{x}^i}$ is the Jacobian to be propagated, and $J_{\mathbf{x}^{i-1}}^{\mathbf{x}^i}$ is the local Jacobian matrix. The Einstein’s notation [35] is used here so that the duplicated index \mathbf{x}^i is summed over. Eq. (1) can be rewritten in the diagram of tensor networks [36] as shown in Fig. 3.

The algorithm to compute the adjoint mode AD can be summarized as follows.

Algorithm 1: Reversible programming AD

Result: `grad(xg)`
 let `iloss` be the index of loss variable in `x`
`y = f(x)`
`yg = GVar(y)`
`grad(yg[iloss]) += 1.0`
`xg = f-1(yg)`

The program first computes the forward pass, and then wrap each output variable with `GVar`. The constructor `GVar` attaches a zero gradient field to a variable. If an input variable is an array, `GVar` will be broadcasted to each array element automatically. The line `grad(yg[iloss]) += 1.0` adds one to the gradient field of loss to initialize a single row of Jacobian as in the first line of Eq. (1). Finally, execute the inverse program f^{-1} , the gradients are stored in the `grad` dataview of output variables. The computation of gradients

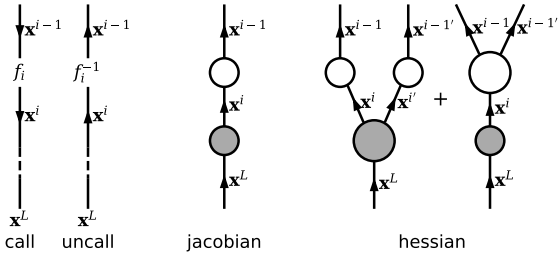


Figure 3: Computational processes in the tensor network diagram, a big circle with three legs represents a Hessian, a small circle with two legs represents a Jacobian. Dangling edges and connected edges stands for unpaired and paired labels respectively in the Einstein's notation. From left to right, the diagrams represent computing, uncomputing, back propagating Jacobians and back propagating Hessians.

```
julia> using NiLang, NiLang.AD

julia> a, b, y = GVar(0.5), GVar(0.6), GVar(0.9)
(GVar(0.5, 0.0), GVar(0.6, 0.0), GVar(0.9, 0.0))

julia> @instr grad(y) += identity(1.0)

julia> @instr y += a * b
GVar(0.6, -0.5)

julia> a, b, y
(GVar(0.5, -0.6), GVar(0.6, -0.5), GVar(1.2, 1.0))

julia> @instr y -= a * b
GVar(0.6, 0.0)

julia> a, b, y
(GVar(0.5, 0.0), GVar(0.6, 0.0), GVar(0.899999, 1.0))
```

Here, $J(\Theta(*)) = J(\Theta(*))^{-1}$, hence consecutively applying them restores gradient fields of variables. The implementation of Algorithm 1 is so short that we present the function definition as follows.

are implemented with multiple dispatch, that is, when an instruction has a `GVar` type in its argument list, it calls a different routine. The same trick is used in the dual number implementation of tangent mode AD [37]. In the following, we exemplify the design by binding the backward rules to instructions $\Theta(*)$ and $\ominus(*)$

```
@i function  $\ominus(*)$ (out!::GVar, x::GVar, y::GVar)
    value(out!) -= value(x) * value(y)
    grad(x) += grad(out!) * value(y)
    grad(y) += value(x) * grad(out!)
end
```

Here, we adopt a convention that only variables ended with `!` will be changed after the function call. Although the backward rule is defined on $\ominus(*)$, the compiler generates the backward rules on $\Theta(*)$ too. This reflects the fact that taking inverse and computing gradients commute to each other [38]. Hence for a general reversible function f , one can bind backward rule on either f or its inverse f^{-1} . We can check the correctness of our definition like follows.

```
@i function (g::Grad)(args...; kwargs...)
    @safe @assert count(x -> x isa Loss, args) == 1
    iloss ← 0
    @routine for i=1:length(args)
        if (tget(args,i) isa Loss, iloss==i)
            iloss += identity(i)
            (~Loss)(tget(args,i))
        end
    end

    g.f(args...; kwargs...)
    GVar.(args)
    grad(tget(args,iloss)) += identity(1.0)
    (~g.f)(args...; kwargs...)

    ~@routine
end
```

The program first checks variables contain exactly one `Loss`, where `Loss` is a reversible wrapper used to mark the loss variable. Then we locates the loss variable as `iloss` and use `~Loss` unwraps the loss variable. After computing the forward pass and backward pass, `~@routine` uncomputes the ancilla `iloss` and returns the location information to the loss variable. Here, `tget(args, i)` returns the i -th element of a tuple. We forbid tuple indexing deliberately in order to avoid possible ambiguity in supporting array indexing.

The overhead of using `GVar` type is negligible thanks to Julia's multiple-dispatch and type inference. Let us consider a simple example that accumulates 1.0 to a target variable x for n times.

[JG: Grammarly here!]

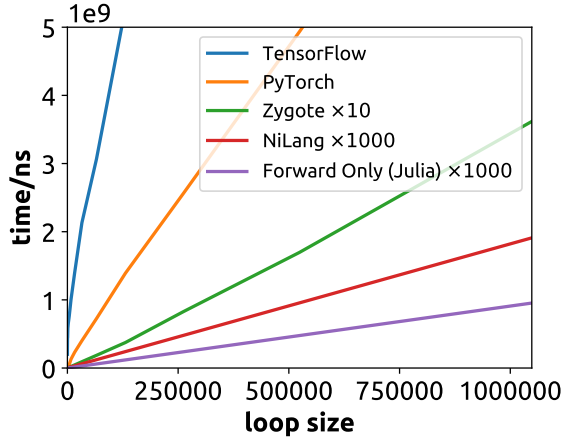


Figure 4: The time to obtain gradient as function of loop size. $\times n$ in legend represents a rescaling of time.

```
julia> using NiLang, NiLang.AD, BenchmarkTools

julia> @i function prog(x, one, n::Int)
    for i=1:n
        x += identity(one)
    end
end

julia> @benchmark prog'(Loss(0.0), 1.0, 10000)
BenchmarkTools.Trial:
  memory estimate:  1.05 KiB
  allocs estimate:  39
  ------------
  minimum time:     35.838 μs (0.00% GC)
  median time:      36.055 μs (0.00% GC)
  mean time:        36.483 μs (0.00% GC)
  maximum time:     185.973 μs (0.00% GC)
  ------------
  samples:          10000
  evals/sample:     1
```

We implement the same function with TensorFlow, PyTorch and Zygote for comparison. The code could be found in our paper’s github repository [39]. Benchmark results on CPU Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz are shown in Fig. 4. One can see that the NiLang implementation is unreasonably fast, it is approximately two times the forward pass written in native Julia code. Reversible programming is not always as fast as its irreversible counterparts. In practical applications, a reversible program may have memory or computation overhead. We will discuss the details of time and space trade off in Sec. V A.

B. Second-order gradient

Second-order gradients can be obtained in two different approaches.

1. Back propagating first-order gradients

Back propagating the first-order gradients is the most widely used approach to obtain the second-order gradients. Suppose the function space is closed under gradient operation, one can obtain higher-order gradients by recursively differentiating lower order gradient functions without defining new backward rules.

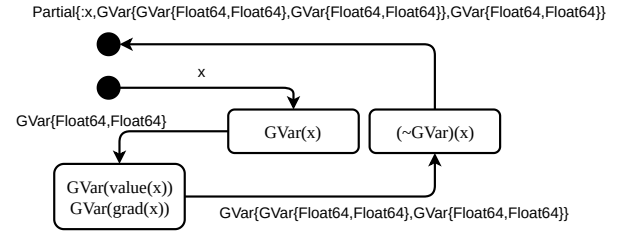


Figure 5: Data flow in obtaining the second-order gradient with the recursive differentiation approach. Annotations on lines are data types used in the computation.

Fig. 5 shows the data flow in the four passes of computing Hessian. The first two passes obtain the gradients. Before entering the third pass, the program wraps each field in GVar with another layer of GVar. Then we pick a variable x_i and add 1 to the gradient field of its gradient $\text{grad}(\text{grad}(x_i))$ in order to compute the i -th row of Hessian. Before entering the final pass, the $\sim\text{GVar}$ is called. We can not unwrap GVar directly because although the values of gradients have been uncomputed to zero, the gradient fields of gradients may be nonzero. Instead, we use $\text{Partial}\{x\}(\text{obj})$ to take field x of an object without erasing memory. By repeating the above procedure for different x_i , one can obtain the full Hessian matrix.

2. Hessian propagation

A probably more efficient approach is back-propagating Hessians directly [40] using the relation

$$H_{\mathbf{x}^{L'}, \mathbf{x}^{L''}}^{\mathbf{x}^L} = \mathbf{0},$$

$$H_{\mathbf{x}^{i-1}, \mathbf{x}^{i-1'}}^{\mathbf{x}^L} = J_{\mathbf{x}^{i-1}}^{\mathbf{x}^L} H_{\mathbf{x}^i, \mathbf{x}^i}^{\mathbf{x}^L} J_{\mathbf{x}^{i-1'}}^{\mathbf{x}^i} + J_{\mathbf{x}^i}^{\mathbf{x}^L} H_{\mathbf{x}^{i-1}, \mathbf{x}^{i-1'}}^{\mathbf{x}^i}. \quad (2)$$

Here, the Hessian tensor $H_{\mathbf{x}^{i-1}, \mathbf{x}^{i-1'}}^{\mathbf{x}^L}$ is rank three, where the top index is often taken as a scalar and omitted. In tensor network diagram, the above equation can be represented as the right panel of Fig. 3. Hessian propagation is a special case of Taylor propagation. With respect to the order of gradients,

Taylor propagation is exponentially more efficient in obtaining higher-order gradients than differentiating lower order gradients recursively. Comparing with operator overloading, source to source automatic differentiation has the advantage of having very limited primitives, exhausted implementation of Hessian propagation is possible. An example to obtain Hessians is provided in Sec. IV B.

C. Gradient on ancilla problem

In this section, we introduced an easily overlooked problem in our reversible AD framework. An ancilla sometimes can carry a nonzero gradient when it is going to be deallocated. As a result, even if an ancilla can be uncomputed rigorously in the original program, its GVar wrapped version is not necessarily safely deallocated. In NiLang, we simply “drop” the gradient field of ancillas instead of raising an error. In the following, we justify our approach by proving the following theorem

Theorem 1. *Deallocating an ancilla with emptied value field and nonzero gradient field does not harm the reversibility of a function.*

Proof. Consider a reversible function $\mathbf{x}^i, b = f_i(\mathbf{x}^{i-1}, a)$, where a and b are the input and output values of an ancilla. Since the ancilla is emptied for any input \mathbf{x}^{i-1} , we have

$$\frac{\partial b}{\partial \mathbf{x}^{i-1}} = \mathbf{0}. \quad (3)$$

Since the gradient fields are derived from the value fields of variables, discarding gradients should not have any effect to the value fields. The rest is to show $\text{grad}(b) \equiv \frac{\partial \mathbf{x}^L}{\partial b}$ does appear in the backward rule of this function. It can be seen from the back-propagation rule

$$\frac{\partial \mathbf{x}^L}{\partial \mathbf{x}^{i-1}} = \frac{\partial \mathbf{x}^L}{\partial \mathbf{x}^i} \frac{\partial \mathbf{x}^i}{\partial \mathbf{x}^{i-1}} + \frac{\partial \mathbf{x}^L}{\partial b} \frac{\partial b}{\partial \mathbf{x}^{i-1}}, \quad (4)$$

where the second term with $\frac{\partial \mathbf{x}^L}{\partial b}$ vanishes naturally. \square

IV. EXAMPLES

A. Computing Fibonacci Numbers

The following is an example that everyone likes, computing Fibonacci number recursively.

```
using NiLang

@i function rfib(out!, n::T) where T
    n1 ← zero(T)
    n2 ← zero(T)
    @routine begin
        n1 += identity(n)
        n1 -= identity(1)
        n2 += identity(n)
        n2 -= identity(2)
    end
    if (value(n) <= 2, ~)
        out! += identity(1)
    else
        rfib(out!, n1)
        rfib(out!, n2)
    end
    ~@routine
end
```

The time complexity of this recursive algorithm is exponential to input n . It is also possible to write a reversible linear time for loop algorithm. A slightly non-trivial task is computing the first Fibonacci number that greater or equal to a certain number z , where a `while` statement is required.

```
@i function rfibn(n!, z)
    @safe @assert n! == 0
    out ← 0
    rfib(out, n!)
    while (out < z, n! != 0)
        ~rfib(out, n!)
        n! += identity(1)
        rfib(out, n!)
    end
    ~rfib(out, n!)
end
```

In this example, the postcondition $n!=0$ in the `while` statement is false before entering the loop, and becomes true in later iterations. In the reverse program, the `while` statement stops at $n==0$. If executed correctly, a user will see the following result.

```
julia> rfib(0, 10)
(55, 10)

julia> rfibn(0, 100)
(12, 100)

julia> (~rfibn)(rfibn(0, 100)...)
(0, 100)
```

This exemplifies how an addition postcondition provided by user can help reversing a control flow without caching

controls.

B. Bessel function

An Bessel function of the first kind of order ν can be computed using Taylor expansion

$$J_\nu(z) = \sum_{n=0}^{\infty} \frac{(z/2)^\nu}{\Gamma(k+1)\Gamma(k+\nu+1)} (-z^2/4)^n \quad (5)$$

where $\Gamma(n) = (n-1)!$ is the Gamma function. One can compute the accumulated item iteratively as $s_n = -\frac{z^2}{4}s_{n-1}$. The irreversible implementation is

```
function besselj( $\nu$ , z; atol=1e-8)
  k = 0
  s = (z/2)^ $\nu$  / factorial( $\nu$ )
  out = s
  while abs(s) > atol
    k += 1
    s *= (-1) / k / (k+ $\nu$ ) * (z/2)^2
    out += s
  end
  out
end
```

This computational process could be diagrammatically represented as a DAG as show in Fig. 6(a). In this diagram, the data is represented as an edge, it connects at most two nodes, one generates this data and one consumes it as an input parameter. Computational graph is more likely a mathematical expression, it can not describe inplace functions or control flows conveniently because it does not have the concept of memory.

In the following, we introduce the reversible implementation and the memory oriented computational graph. Intuitively, one needs an increasing size of tape to cache the intermediate state, since one can not release state s_{n-1} directly after computing s_n [17]. To reduce the memory allocation without increasing the time complexity of program, we introduce the following reversible approximate multiplier

```
1 @i @inline function imul(out!, x, anc!)
2   anc! += out! * x
3   out! -= anc! / x
4   SWAP(out!, anc!)
5 end
```

Here, the definition of SWAP instruction can be found in Appendix B, $\text{anc!} \approx 0$ is a *dirty ancilla*. Line 2 computes the result and accumulates it to the dirty ancilla, we get an approximately correct output in anc! . Line 3 “uncomputes” out! approximately by using the information stored in anc! , leaving a dirty zero state in register out! . Line 4

swaps the contents in out! and anc! . Finally, we have an approximately correct output and a dirtier ancilla. With this multiplier, we implementation J_ν as follows.

```
using NiLang, NiLang.AD

@i function ibesselj(out!,  $\nu$ , z; atol=1e-8)
  k ← 0
  fact_nu ← zero( $\nu$ )
  halfz ← zero(z)
  halfz_power_nu ← zero(z)
  halfz_power_2 ← zero(z)
  out_anc ← zero(z)
  anc1 ← zero(z)
  anc2 ← zero(z)
  anc3 ← zero(z)
  anc4 ← zero(z)
  anc5 ← zero(z)

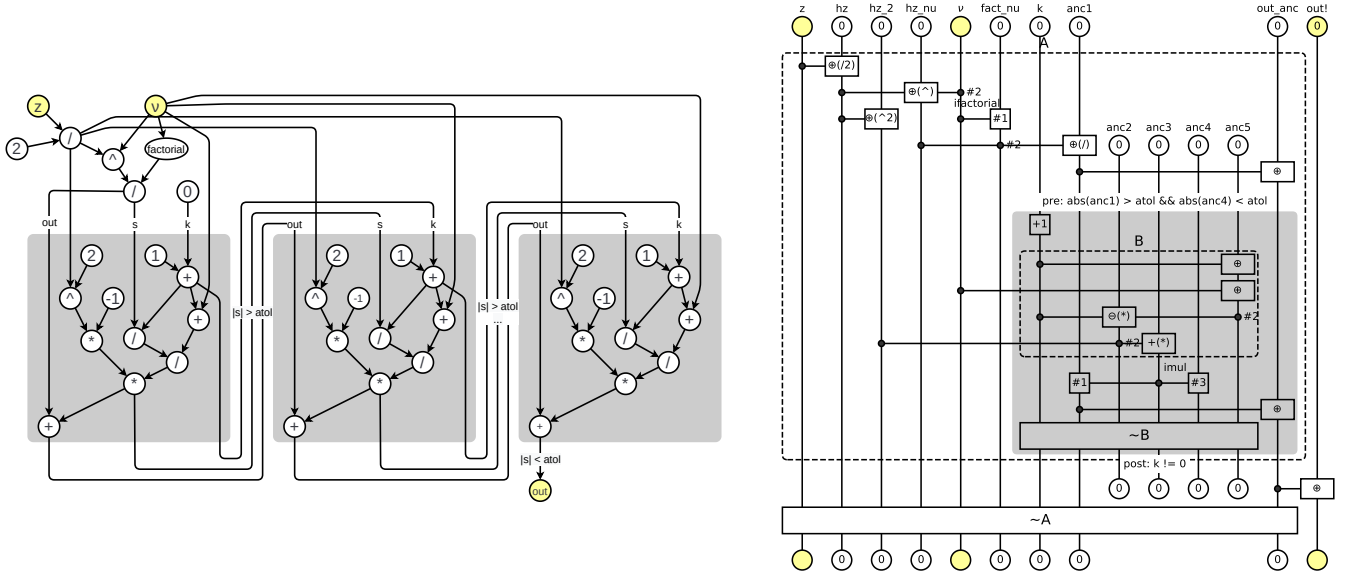
  @routine begin
    halfz += z / 2
    halfz_power_nu += halfz ^  $\nu$ 
    halfz_power_2 += halfz ^ 2
    ifactorial(fact_nu,  $\nu$ )

    anc1 += halfz_power_nu/fact_nu
    out_anc += identity(anc1)
    while (abs(unwrap(anc1)) > atol &&
           abs(unwrap(anc4)) < atol, k!=0)
      k += identity(1)
      @routine begin
        anc5 += identity(k)
        anc5 += identity( $\nu$ )
        anc2 -= k * anc5
        anc3 += halfz_power_2 / anc2
      end
      imul(anc1, anc3, anc4)
      out_anc += identity(anc1)
    ~@routine
  end

  out! += identity(out_anc)
  ~@routine
end

@i function ifactorial(out!, n)
  out! += identity(1)
  for i=1:n
    MULINT(out!, i)
  end
end
```

Here, only a constant number of ancillas are used in this implementation, while the algorithm complexity does not increase comparing to its irreversible counterpart. ancilla anc4 plays the role of *dirty ancilla* in multiplication, it is uncomputed rigorously in the uncomputing stage. The reason why the “approximate uncomputing” trick works here lies in the fact that from the mathematic perspective the state in n th step $\{s_n, z\}$ contains the same amount of information as the state in the $n-1$ th step $\{s_{n-1}, z\}$ except some special points, it is highly possible to find an equation to uncompute the previous state from the current state. This trick can be



(a) Where a vertex (circle) is an operation and a directed edge is a variable. The gray region is the body of while loop. The while loop is unrolled.

(b) Vertical lines are variables, boxes and dots connected by parallel lines are operations. A dot represents using a variable as control while a box represents mutating a variable. The gray box represents the body of a while loop, where a precondition and postcondition are positioned on top and bottom of this box.

Figure 6: The comparison between (a) traditional and (b) memory oriented computational graph of the first kind Bessel function.

used extensively in many other application. It mitigated the artificial irreversibility brought by the number system that we have adopt at the cost of precision.

This reversible program can be diagrammatically represented as a memory oriented computational graph as show in Fig. 6 (b). The key difference is we adopt a variable centric view, while a variable can be used by multiple operations. If a function uses a variable but does not change the contents of it, we call this variable a control parameter in this function. It is represented as a dot in the graph. Otherwise if it is changed, it is represented as a square. This diagram can be used to analyse uncomputable variables. In this example routine “B” changes anc2, anc3 and anc5. while the following operation imul does not change these variables. Hence we can apply the inverse routine ~B to safely restore contents anc2, anc3 and anc5, this is what we called compute-copy-uncompute paradigm.

To obtain gradients, one can wrap the variable out!! with Loss and feed it into ibesselj'

Here, `ibesselj'` is a callable instance of type `Grad{typeof(ibesselj)}`. Its implementation is shown in Sec. III A. This function itself is reversible and differentiable, one can back-propagate this function to obtain Hessians as introduced in Sec. III B 1. In NiLang, it is implemented as `hessian_repeat`.

```
julia> hessian_repeat(ibesselj, (Loss(out!), 2, x))
3×3 Array{Float64,2}:
 0.0  0.0  0.0
 0.0  0.0  0.0
 0.0  0.0 -0.27505
```

```
julia> out!, x = 0.0, 3.0
(0.0, 3.0)

julia> ibesselj'(Loss(out!), 2, x)
(Loss(GVar(0.0, 1.0)), 2, GVar(3.0, 0.0149981304))
```

To obtain Hessians, we can also use the Hessian propagation approach as introduced in Sec. III B 2.

```
julia> ibesselj''(Loss(out!), 2, x)
(Loss(BeijingRing{Float64}(0.0, 1.0, 1)), 2,
BeijingRing{Float64}(3.0, 0.014998134978750133, 2))

julia> collect_hessian()
2×2 Array{Float64,2}:
 0.0  0.0
 0.0 -0.27505
```

`ibesselj''` computes the second-order gradients. It wraps variables with type `BeijingRing` [41] in the backward pass. `BeijingRing` records Jacobians and Hessians for a variable, where Hessians are stored in a global storage. Whenever an n -th variable or ancilla is created, we push a ring of size $2n - 1$ to a global tape. Whenever an ancilla is deallocated, we pop a ring from the top. The n -th ring stores Hessian elements $H_{i \leq n, n}$ and $H_{n, i < n}$. The final result can be collected by calling `collect_hessian()`, which will read out the Hessian matrix that stored in the global storage.

C. Unitary Matrices

A unitary matrices features uniform eigenvalues and reversibility. It is widely used as an approach to ease the gradient exploding and vanishing problem [42–44] and the memory wall problem [9]. One of the simplest way to parametrize a unitary matrix is representing a unitary matrix as a product of two-level unitary operations [44]. A real unitary matrix of size N can be parametrized compactly by $N(N - 1)/2$ rotation operations [45]

$$\text{ROT}(a!, b!, \theta) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} a! \\ b! \end{pmatrix}, \quad (6)$$

where θ is the rotation angle, $a!$ and $b!$ are target registers.

```
using NiLang, NiLang.AD

@i function umm!(x!, θ)
    @safe @assert length(θ) ==
        length(x!)*(length(x!)-1)/2
    k ← 0
    for j=1:length(x!)
        for i=length(x!)-1:-1:j
            k += identity(1)
            ROT(x![i], x![i+1], θ[k])
        end
    end
    k → length(θ)
end
```

Here, the ancilla k is deallocated manually by specifying its value, because we know the loop size is $N(N - 1)/2$. We define the test functions in order to check gradients.

```
julia> @i function isum(out!, x::AbstractArray)
    for i=1:length(x)
        out! += identity(x[i])
    end
end

julia> @i function test!(out!, x::Vector, θ::Vector)
    umm!(x!, θ)
    isum(out!, x!)
end

julia> out, x, θ = Loss(0.0), randn(4), randn(6);

julia> @instr test!(out, x, θ)

julia> x
4-element Array{GVar{Float64,Float64},1}:
 GVar(1.220182125326287, 0.14540743042341095)
 GVar(2.1288634811475937, -1.3749962375499805)
 GVar(1.2696579252569677, 1.42868739498625)
 GVar(0.1083891125379283, 0.2170123344615735)

julia> @instr (~test!')(out, x, θ)

julia> x
4-element Array{Float64,1}:
 1.220182125326287
 2.128863481147593
 1.269657925256967
 0.10838911253792821
```

In the above testing code, `test'` attaches a gradient field to each element of x . `~test'` is the inverse program that erase the gradient fields. Notably, this reversible implementation costs zero memory allocation although it changes the target variables inplace.

D. QR decomposition

Let's consider a naive implementation of QR decomposition from scratch. We admit this implementation is just a proof of principle which does not even consider reorthogonalization.

```

using NiLang, NiLang.AD

@i function qr(Q, R, A::Matrix{T}) where T
    anc_norm ← zero(T)
    anc_dot ← zeros(T, size(A,2))
    ri ← zeros(T, size(A,1))
    for col = 1:size(A, 1)
        ri .+= identity.(A[:,col])
        for precol = 1:col-1
            dot(anc_dot[precol], Q[:,precol], ri)
            R[precol,col] +=
                identity(anc_dot[precol])
            for row = 1:size(Q,1)
                ri[row] -=
                    anc_dot[precol] * Q[row, precol]
            end
        end
        norm2(anc_norm, ri)

        R[col, col] += anc_norm^0.5
        for row = 1:size(Q,1)
            Q[row,col] += ri[row] / R[col, col]
        end
    end

    ~begin
        ri .+= identity.(A[:,col])
        for precol = 1:col-1
            dot(anc_dot[precol], Q[:,precol], ri)
            for row = 1:size(Q,1)
                ri[row] -= anc_dot[precol] *
                    Q[row, precol]
            end
        end
        norm2(anc_norm, ri)
    end
end
end

```

Here, in order to avoid frequent uncomputing, we allocate ancillas `ri` and `anc_dot` as vectors. The expression in `~` is used to uncompute `ri`, `anc_dot` and `anc_norm`. `dot` and `norm2` are reversible functions to compute dot product and vector norm. Their implementations could be found in Appendix D. One can easily check the correctness of the gradient function

```

julia> A, q, r = randn(4,4), zero(A), zero(A);
julia> @i function test1(out, q, r, A)
    qr(q, r, A)
    isum(out, q)
end
julia> check_grad(test1, (Loss(0.0), q, r, A))
true

```

Here, the loss function `test1` is defined as the sum of the output unitary matrix `q`. The `check_grad` function is a gradient checker function defined in module `NiLang.AD`.

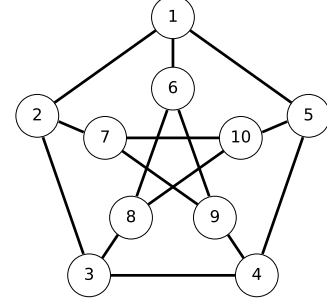


Figure 7: The Petersen graph with 10 vertices and 15 edges.

E. Solving a graph embedding problem

Graph embedding has an interesting application of finding a representation for an order parameter [46] in condensed matter physics. Ref. 46 considers a problem of finding the minimum Euclidean space dimension k that a Petersen graph can fit into, with extra requirements that the distance between a pair of connected vertices has the same value l_1 , the distance between a pair of disconnected vertices has the same value l_2 and $l_2 > l_1$. The Petersen graph is 10 vertices graph as shown in Fig. 7. Let's denote the set of connected and disconnected vertex pairs as L_1 and L_2 respectively. This problem can be variationally solved by differential programming by designing the following loss.

$$\mathcal{L} = \text{Var}(\text{dist}(L_1)) + \text{Var}(\text{dist}(L_2)) + \exp(\text{relu}(\text{E}(\text{dist}(L_1)) - \text{E}(\text{dist}(L_2)) + 0.1))) - 1 \quad (7)$$

The first line is a summation of distance variances in two sets of vertex pairs. The second line is used to guarantee $l_2 > l_1$.


```

using NiLang, NiLang.AD

@i function sqdistance(dist!, x1::AbstractVector{T},
    x2::AbstractVector) where T
    @inbounds for i=1:length(x1)
        x1[i] -= identity(x2[i])
        dist! += x1[i] ^ 2
        x1[i] += identity(x2[i])
    end
end

# bonds of a petersen graph
const L1 = [(1, 6), (2, 7), (3, 8), (4, 9), (5, 10),
    (1, 2), (2, 3), (3, 4), (4, 5), (1, 5), (6, 8),
    (8, 10), (7, 10), (7, 9), (6, 9)]

# disconnected bonds of a petersen graph
const L2 = [(1, 3), (1, 4), (1, 7), (1, 8), (1, 9),
    (1, 10), (2, 4), (2, 5), (2, 6), (2, 8), (2, 9),
    (2, 10), (3, 5), (3, 6), (3, 7), (3, 9), (3, 10),
    (4, 6), (4, 7), (4, 8), (4, 10), (5, 6), (5, 7),
    (5, 8), (5, 9), (6, 7), (6, 10), (7, 8), (8, 9),
    (9, 10)]

@i function embedding_loss(out!::T, x) where T
    v1 ← zero(T)
    m1 ← zero(T)
    v2 ← zero(T)
    m2 ← zero(T)
    diff ← zero(T)
    d1 ← zeros(T, length(L1))
    d2 ← zeros(T, length(L2))
    @routine begin
        for i=1:length(L1)
            sqdistance(d1[i],
                x[:,L1[i][1]],x[:,L1[i][2]])
        end
        for i=1:length(L2)
            sqdistance(d2[i],
                x[:,L2[i][1]],x[:,L2[i][2]])
        end
        var_and_mean_sq(v1, m1, d1)
        var_and_mean_sq(v2, m2, d2)
        m1 -= identity(m2)
        m1 += identity(0.1)
    end
    out! += identity(v1)
    out! += identity(v2)
    if (m1 > 0, ~)
        # to ensure mean(v2) > mean(v1)
        # if mean(v1)+0.1 - mean(v2) > 0, punish it.
        out! += exp(m1)
        out! -= identity(1)
    end
    ~@routine
end

```

One can access the gradient by typing

```
iloss'(Loss(0.0), randn(5, 10))
```

We repeat the training for each dimension k from 1 to 9 and search for possible solutions by variationally optimizing the positions of vertices. In each training, we fixed two of the vertices and train the rest, otherwise the program will find the trivial solution with overlapped vertices. The training employs the Adam optimizer with a learning rate 0.01 [47], a 10000 steps training costs less than 1 second on a laptop. For $k = 5$, we can get a loss close to machine precision with high probability, while for $k < 5$, the loss is always much higher than 0. From the solution, it is easy to verify $l_2/l_1 = \sqrt{2}$.

V. DISCUSSION AND OUTLOOK

In this paper, we show a program on an reversible Turing machine can be differentiated to any order reliably and efficiently without sophisticated designs to memorize computational graph and intermediate states. We introduce a reversible Julia eDSL NiLang that implements a reversible AD. In a reversible programming language, we proposed to use “approximate uncomputing” trick to avoid the overhead of a reversible program in many practical cases.

In the following, we discussed some practical issues about reversible programming, and several future directions to go.

A. Time Space Tradeoff

In history, there has been many other interesting designs of reversible languages. However, current popular programming languages are all irreversible. In the simplest g-segment trade off scheme [48, 49], a RTM model has either a space overhead that proportional to computing time T or a computational overhead that sometimes can be exponential comparing with a irreversible counter part. The tradeoff between space and time is a crucial issue in the theory of RTM. In the following, we try to convince the readers that the overhead of reversible computing is not as terrible as people thought.

The overhead of reversing a program is bounded the checkpointing [50] strategy used in a traditional machine learning package that memorizes every inputs of primitives because similar strategy can also be used in reversible programming. [17] Reversible programming simply provides more alternatives to reduce the overhead. For example, the overhead in many iterative algorithms can often be removed with “arithmetic uncomputing” trick without sacrificing reversibility as shown in the `ibessel.j` example in Sec. IV B.

Clever compiling can also be used to remove most overheads. Often, when we define a new reversible function, we allocate some ancillas at the beginning of the function and deallocate them through uncomputing at the end. The overhead comes from the uncomputing, in the worst case, the time used for uncomputing can be the same as the forward pass. In a hierarchical design, uncomputing can appear in every layer of abstraction. To quantify the overhead of uncomputing, we introducing the concept

Definition 1 (program granularity). The log ratio between

the execution time of a reversible program and its irreversible counter part

The computing time increases exponentially as the granularity increases. A cleverer compilation of a program can reduce the granularity by merging the uncomputing statements to avoid repeated efforts.

At last, making reversible programming an eDSL rather than an independent language allows flexible choices between reversibility and computational overhead. For example, in order to deallocate the gradient memory in a reversible language one has to uncompute the whole process of obtaining this gradient. In our eDSL, we can just deallocate the memory irreversibly, i.e. trade energy with time.

B. Instructions and Hardware

So far, our eDSL is compiled to Julia. In the future, it can be compiled to reversible instructions [51] and executed on a reversible device. However, arithmetic instructions should be redesigned to support better reversible programs. The major obstacle to exact reversibility programming is current floating point adders and multipliers used in our computing devices are not exactly reversible. There are proposals of reversible floating point adders and multipliers [52–55], however these designs with allocate garbage bits in each operation. Alternatives include fixed point numbers [56] and logarithmic numbers [57, 58], where logarithmic number system is reversible under $\ast =$ and $/ =$. With these infrastructures, a reversible program can be executed without suffering from the rounding error.

Reversible instructions can be executed on an energy efficient reversible hardware. In the introduction, we mentioned several reversible hardware. A reversible hardware can be those supporting reversible gates such as the Toffoli gate and the Fredkin gate, or like an adiabatic CMOS with the ability to recover signal energy. The latter is known as the generalized reversible computing. [59, 60] In the near future, there might be energy efficient artificial intelligence (AI) chips as coprocessors that our eDSL can compile to. Since reversible computing is mainly driven by quantum computing in recent years. In the following, we comment briefly on quantum devices.

1. Quantum Computers

Building a universal quantum computer [61] is difficult. The difficulty lies in the fact that it is extremely hard to protect a quantum state. Unlike a classical state, an quantum state is can not be cloned, meanwhile, it losses information by interacting with the environment, or decoherence. Classical reversible computing does not enjoy the quantum advantage, nor the quantum disadvantages of non-cloning and decoherence. It is technically more smooth to have a reversible computing device to bridge the gap between classical devices and universal quantum computing devices. By introducing

entanglement little by little, we can accelerate some basic components in reversible computing. For example, quantum Fourier transformation provides an interesting alternative to the reversible adders and multipliers by introducing the CPHASE quantum gate. [62] The development of reversible compiling theory can be benefit quantum compiling directly.

C. Outlook

The reversible eDSL NiLang can be used to solve many existing scientific computing problems. First of all, it can be used to generate AD rules for existing machine learning packages like Zygote. For example, one can use NiLang to generate backward rules for singular value decomposition and eigenvalue decomposition functions that extensively used in scientific computing [63, 64]. Although their backward rules [65–67] have been driven in recent years, these backward rules can not handle degenerate eigenvalues properly. Hopefully, the automatically generated backward rules do not have such problems.

Secondly, we can use it to overcome the memory wall problem in some applications. NiLang provides a systematic time-space trade off scheme through uncomputing. A successful related example is the memory efficient domain-specific AD engine in quantum simulator Yao [9]. This domain-specific AD engine is written in a reversible style and solved the memory bottleneck in variational quantum simulations. It also gives hitherto the best performance in differentiating quantum circuit parameters. Similarly, we can write memory efficient normalizing flow [68] with NiLang. Normalizing flow is a successful class of generative model in both computer vision [69] and quantum physics [70, 71], where its building block bijector is reversible. We can use similar idea to differentiate reversible integrators [72, 73]. With reversible integrators, it should be possible to rewrite the control system in robotics [74] in a reversible style, where scalars are first class citizen rather than tensors. Writing a control program reversibly should boost the training performance a lot.

Thirdly, reversibility is a resource for training. For those who are interested in non-gradient based training. In Appendix C, we provide a self-consistency training strategy for reversible programs.

Latstly, the reversible IR is a good starting point to study quantum compiling. Most quantum programming language preassumes a classical coprocessor and use classical control flows [75] in universal quantum computing. However, we believe reversible control flows are also very important to a universal quantum computer.

To solve the above problems better, NiLang can be improved from multiple perspectives. Like we need a more efficient fixed point or log number system to avoid rounding errors. Currently we simulate reversible arithmetics with the julia fixed point numebr package. [56] Then we should optimize the compiling to decreases granularity of a program and reduces uncomputing overheads. There are also some known issues to be solved like the type inference problem, we have listed some of them on Github. These improvements

needs participation of people from multiple fields.

VI. ACKNOWLEDGMENTS

Jin-Guo Liu thank Lei Wang for motivating the project with possible applications reversible integrator, normalizing flow and neural ODE. Xiu-Zhe Luo for discussion on the implementation details of source to source automatic differentiation, Shuo-Hui Li for helpful discussion on differential

geometry. Damian Steiger for telling me the come from joke. Tong Liu and An-Qi Chen for helpful discussion on quantum adders and multipliers. Chris Rackauckas for helpful discussion on reversible integrator. Mike Innes for reviewing the comments about Zygote. Jun Takahashi for discussion about the graph embedding problem. Simon Byrne and Chen Zhao for helpful discussion on floating point and logarithmic numbers. The authors are supported by the National Natural Science Foundation of China under the Grant No. 11774398, the Strategic Priority Research Program of Chinese Academy of Sciences Grant No. XDB28000000.

-
- [1] L. Hascoet and V. Pascual, *ACM Transactions on Mathematical Software (TOMS)* **39**, 20 (2013).
 - [2] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, in *NIPS Autodiff Workshop* (2017).
 - [3] M. Innes, E. Saba, K. Fischer, D. Gandhi, M. C. Rudilosso, N. M. Joy, T. Karmali, A. Pal, and V. Shah, “Fashionable modelling with flux,” (2018), [arXiv:1811.01457 \[cs.PL\]](#).
 - [4] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” (2015), software available from tensorflow.org.
 - [5] M. Innes, “Don’t unroll adjoint: Differentiating ssa-form programs,” (2018), [arXiv:1810.07951 \[cs.PL\]](#).
 - [6] M. Innes, A. Edelman, K. Fischer, C. Rackauckas, E. Saba, V. B. Shah, and W. Tebbutt, *CoRR* **abs/1907.07587** (2019), [arXiv:1907.07587](#).
 - [7] Z. C. Lipton, J. Berkowitz, and C. Elkan, “A critical review of recurrent neural networks for sequence learning,” (2015), [arXiv:1506.00019 \[cs.LG\]](#).
 - [8] K. He, X. Zhang, S. Ren, and J. Sun, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 10.1109/cvpr.2016.90.
 - [9] X.-Z. Luo, J.-G. Liu, P. Zhang, and L. Wang, “Yao.jl: Extensible, efficient framework for quantum algorithm design,” (2019), [arXiv:1912.10877 \[quant-ph\]](#).
 - [10] M. MacKay, P. Vicol, J. Ba, and R. B. Grosse, in *Advances in Neural Information Processing Systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018) pp. 9029–9040.
 - [11] J. Behrmann, D. Duvenaud, and J. Jacobsen, *CoRR* **abs/1811.00995** (2018), [arXiv:1811.00995](#).
 - [12] D. Maclaurin, D. Duvenaud, and R. Adams, in *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 37, edited by F. Bach and D. Blei (PMLR, Lille, France, 2015) pp. 2113–2122.
 - [13] A. N. Gomez, M. Ren, R. Urtasun, and R. B. Grosse, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017) pp. 2214–2224.
 - [14] J.-H. Jacobsen, A. W. Smeulders, and E. Oyallon, in *International Conference on Learning Representations* (2018).
 - [15] J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman, *arXiv preprint arXiv:1209.5145* (2012).
 - [16] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *SIAM Review* **59**, 65–98 (2017).
 - [17] K. S. Perumalla, *Introduction to reversible computing* (Chapman and Hall/CRC, 2013).
 - [18] M. P. Frank, *IEEE Spectrum* **54**, 32–37 (2017).
 - [19] C. Lutz, “Janus: a time-reversible language,” (1986), *Letter to R. Landauer*.
 - [20] M. P. Frank, *The R programming language and compiler*, Tech. Rep. (MIT Reversible Computing Project Memo, 1997).
 - [21] I. Lanese, N. Nishida, A. Palacios, and G. Vidal, *Journal of Logical and Algebraic Methods in Programming* **100**, 71–97 (2018).
 - [22] T. Haulund, “Design and implementation of a reversible object-oriented programming language,” (2017), [arXiv:1707.07845 \[cs.PL\]](#).
 - [23] M. P. Frank and T. F. Knight Jr, *Reversibility for efficient computing*, Ph.D. thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and ... (1999).
 - [24] J. G. Koller and W. C. Athas, in *Workshop on Physics and Computation* (1992) pp. 267–270.
 - [25] R. C. Merkle, R. A. Freitas, T. Hogg, T. E. Moore, M. S. Moses, and J. Ryley, *Journal of Mechanisms and Robotics* **10** (2018), 10.1115/1.4041209.
 - [26] K. Likharev, *IEEE Transactions on Magnetics* **13**, 242 (1977).
 - [27] V. K. Semenov, G. V. Danilov, and D. V. Averin, *IEEE Transactions on Applied Superconductivity* **13**, 938 (2003).
 - [28] R. Landauer, *IBM journal of research and development* **5**, 183 (1961).
 - [29] E. P. DeBenedictis, J. K. Mee, and M. P. Frank, *Computer* **50**, 76 (2017).
 - [30] D. R. Jefferson, *ACM Transactions on Programming Languages and Systems (TOPLAS)* **7**, 404 (1985).
 - [31] B. Boothe, *ACM SIGPLAN Notices* **35**, 299 (2000).
 - [32] G. H. Golub and C. F. Van Loan, *Matrix computations*, Vol. 3 (JHU press, 2012).
 - [33] C. H. Bennett (1973).
 - [34] “MLStyle.jl,” <https://github.com/thautwarm/MLStyle.jl>.
 - [35] Wikipedia contributors, “Einstein notation — Wikipedia, the free encyclopedia,” (2020), [Online; accessed 20-February-2020].
 - [36] R. Orús, *Annals of Physics* **349**, 117–158 (2014).

- [37] J. Revels, M. Lubin, and T. Papamarkou, “Forward-mode automatic differentiation in julia,” (2016), [arXiv:1607.07892 \[cs.MS\]](#).
- [38] A. McInerney, *First steps in differential geometry* (Springer, 2015).
- [39] “Paper’s Github Repository,” <https://github.com/GiggleLiu/nilangpaper/tree/master/codes>.
- [40] J. Martens, I. Sutskever, and K. Swersky, “Estimating the hessian by back-propagating curvature,” (2012), [arXiv:1206.6464 \[cs.LG\]](#).
- [41] When people ask for the location in Beijing, they will start by asking which ring it is? We use the similar approach to locate the elements of Hessian matrix.
- [42] M. Arjovsky, A. Shah, and Y. Bengio, *CoRR abs/1511.06464* (2015), [arXiv:1511.06464](#).
- [43] S. Wisdom, T. Powers, J. R. Hershey, J. L. Roux, and L. Atlas, “Full-capacity unitary recurrent neural networks,” (2016), [arXiv:1611.00035 \[stat.ML\]](#).
- [44] L. Jing, Y. Shen, T. Dubcek, J. Peurifoy, S. A. Skirlo, M. Tegmark, and M. Soljacic, *CoRR abs/1612.05231* (2016), [arXiv:1612.05231](#).
- [45] C.-K. LI, R. ROBERTS, and X. YIN, *International Journal of Quantum Information* **11**, 1350015 (2013).
- [46] J. Takahashi and A. W. Sandvik, “Valence-bond solids, vestigial order, and emergent $so(5)$ symmetry in a two-dimensional quantum magnet,” (2020), [arXiv:2001.10045 \[cond-mat.str-el\]](#).
- [47] D. P. Kingma and J. Ba, [arXiv:1412.6980](#).
- [48] C. H. Bennett, *SIAM Journal on Computing* **18**, 766 (1989).
- [49] R. Y. Levine and A. T. Sherman, *SIAM Journal on Computing* **19**, 673 (1990).
- [50] T. Chen, B. Xu, C. Zhang, and C. Guestrin, *CoRR abs/1604.06174* (2016), [arXiv:1604.06174](#).
- [51] C. J. Vieri, *Reversible Computer Engineering and Architecture, Ph.D. thesis*, Cambridge, MA, USA (1999), aAI0800892.
- [52] M. Nachtigal, H. Thapliyal, and N. Ranganathan, in *10th IEEE International Conference on Nanotechnology* (2010) pp. 233–237.
- [53] M. Nachtigal, H. Thapliyal, and N. Ranganathan, in *2011 11th IEEE International Conference on Nanotechnology* (2011) pp. 451–456.
- [54] T. D. Nguyen and R. V. Meter, “A space-efficient design for reversible floating point adder in quantum computing,” (2013), [arXiv:1306.3760 \[quant-ph\]](#).
- [55] T. Häner, M. Soeken, M. Roetteler, and K. M. Svore, “Quantum circuits for floating-point arithmetic,” (2018), [arXiv:1807.02023 \[quant-ph\]](#).
- [56] “FixedPointNumbers.jl,” <https://github.com/JuliaMath/FixedPointNumbers.jl>.
- [57] F. J. Taylor, R. Gill, J. Joseph, and J. Radke, *IEEE Transactions on Computers* **37**, 190 (1988).
- [58] “LogarithmicNumbers.jl,” <https://github.com/cjdoris/LogarithmicNumbers.jl>.
- [59] M. P. Frank, in *35th International Symposium on Multiple-Valued Logic (ISMVL’05)* (2005) pp. 168–185.
- [60] M. P. Frank, in *Reversible Computation*, edited by I. Phillips and H. Rahaman (Springer International Publishing, Cham, 2017) pp. 19–34.
- [61] M. A. Nielsen and I. Chuang, “Quantum computation and quantum information,” (2002).
- [62] L. Ruiz-Perez and J. C. Garcia-Escartin, *Quantum Information Processing* **16** (2017), [10.1007/s11128-017-1603-1](#).
- [63] J.-G. L. Hao Xie and L. Wang, [arXiv:2001.04121](#).
- [64] H.-J. Liao, J.-G. Liu, L. Wang, and T. Xiang, *Physical Review X* **9** (2019), [10.1103/physrevx.9.031041](#).
- [65] M. Seeger, A. Hetzel, Z. Dai, E. Meissner, and N. D. Lawrence, “Auto-differentiating linear algebra,” (2017), [arXiv:1710.08717 \[cs.MS\]](#).
- [66] Z.-Q. Wan and S.-X. Zhang, “Automatic differentiation for complex valued svd,” (2019), [arXiv:1909.02659 \[math.NA\]](#).
- [67] C. Hubig, “Use and implementation of autodifferentiation in tensor network methods with complex scalars,” (2019), [arXiv:1907.13422 \[cond-mat.str-el\]](#).
- [68] I. Kobyzev, S. Prince, and M. A. Brubaker, “Normalizing flows: An introduction and review of current methods,” (2019), [arXiv:1908.09257 \[stat.ML\]](#).
- [69] D. P. Kingma and P. Dhariwal, in *Advances in Neural Information Processing Systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018) pp. 10215–10224.
- [70] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using real nvp,” (2016), [arXiv:1605.08803 \[cs.LG\]](#).
- [71] S.-H. Li and L. Wang, *Physical Review Letters* **121** (2018), [10.1103/physrevlett.121.260601](#).
- [72] P. Hut, J. Makino, and S. McMillan, *The Astrophysical Journal* **443**, L93 (1995).
- [73] D. N. Laikov, *Theoretical Chemistry Accounts* **137** (2018), [10.1007/s00214-018-2344-7](#).
- [74] M. Gifftthaler, M. Neunert, M. Stäuble, M. Frigerio, C. Semini, and J. Buchli, *Advanced Robotics* **31**, 1225–1237 (2017).
- [75] K. Svore, M. Roetteler, A. Geller, M. Troyer, J. Azariah, C. Granade, B. Heim, V. Kliuchnikov, M. Mykhailova, and A. Paz, *Proceedings of the Real World Domain Specific Languages Workshop 2018 on - RWDSL2018* (2018), [10.1145/3183895.3183901](#).
- [76] M. Bender, P.-H. Heenen, and P.-G. Reinhard, *Rev. Mod. Phys.* **75**, 121 (2003).
- [77] T. Besard, C. Foket, and B. D. Sutter, *CoRR abs/1712.03112* (2017), [arXiv:1712.03112](#).

Appendix A: NiLang Grammar

To define a reversible function one can use “@i” plus a normal function definition like bellow

```

"""
docstring...
"""
@i function f(args..., kwargs...) where {...}
    <stmts>
end

```

where the definition of “<stmts>” are shown in the grammar on the next column. The following is a list of terminologies used in the definition of grammar

- *ident*, symbols
- *num*, numbers
- ϵ , empty statement
- *JuliaExpr*, native Julia expression
- $[]$, zero or one repetitions.

Here, all *JuliaExpr* should be pure, otherwise the reversibility is not guaranteed. Dataview is a view of a data, it can be a bijective mapping of an object, an item of an array or a field of an object.

```

<Stmts> ::=  $\epsilon$ 
          | <Stmt>
          | <Stmts> <Stmt>
<Stmt> ::= <BlockStmt>
          | <IfStmt>
          | <WhileStmt>
          | <ForStmt>
          | <InstrStmt>
          | <RevStmt>
          | <AncillaStmt>
          | <TypecastStmt>
          | <@routine> <Stmt>
          | <@safe> JuliaExpr
          | <CallStmt>
<BlockStmt> ::= begin <Stmts> end
<RevCond> ::= ( JuliaExpr , JuliaExpr )
<IfStmt> ::= if <RevCond> <Stmts> [else <Stmts>] end
<WhileStmt> ::= while <RevCond> <Stmts> end
<Range> ::= JuliaExpr : JuliaExpr [: JuliaExpr]
<ForStmt> ::= for ident = <Range> <Stmts> end
<KwArg> ::= ident = JuliaExpr
<KwArgs> ::= [<KwArgs> , ] <KwArg>
<CallStmt> ::= JuliaExpr ( [<DataViews>] [; <KwArgs>] )
<Constant> ::= num |  $\pi$ 
<InstrBinOp> ::= += | -= |  $\forall$ =
<InstrTrailer> ::= [.] ( [<DataViews>] )
<InstrStmt> ::= <DataView> <InstrBinOp> ident [<InstrTrailer>]
<RevStmt> ::= ~ <Stmt>
<AncillaStmt> ::= ident  $\leftarrow$  JuliaExpr
<TypecastStmt> ::= ( JuliaExpr => JuliaExpr ) ( ident )
<@routine> ::= @routine ident <Stmt>
<@safe> ::= @safe JuliaExpr
<DataViews> ::=  $\epsilon$ 
               | <DataView>
               | <DataViews> , <DataView>
               | <DataViews> , <DataView> ...
<DataView> ::= <DataView> [ JuliaExpr ]
               | <DataView> . ident
               | JuliaExpr ( <DataView> )
               | <DataView> '
               | - <DataView>
               | <Constant>
               | ident

```

Appendix B: Instructions and Backward Rules

The translation of instructions to Julia functions
The list of instructions implemented in NiLang

instruction	translated	symbol
$y += f(args...)$	$\text{PlusEq}(f)(args...)$	\oplus
$y -= f(args...)$	$\text{MinusEq}(f)(args...)$	\ominus
$y \vee = f(args...)$	$\text{XorEq}(f)(args...)$	\odot

Table II: Instructions and their compilation in NiLang.

instruction	output
$\text{SWAP}(a, b)$	b, a
$\text{ROT}(a, b, \theta)$	$a \cos \theta - b \sin \theta, b \cos \theta + a \sin \theta, \theta$
$\text{IROT}(a, b, \theta)$	$a \cos \theta + b \sin \theta, b \cos \theta - a \sin \theta, \theta$
$\text{MULINT}(a, b)$	$a * b, b$
$\text{DIVINT}(a, b)$	$a/b, b$
$y += a^b$	$y + a^b, a, b$
$y += \exp(x)$	$y + e^x, x$
$y += \log(x)$	$y + \log x, x$
$y += \sin(x)$	$y + \sin x, x$
$y += \cos(x)$	$y + \cos x, x$
$y += \text{abs}(x)$	$y + x , x$
$\text{NEG}(y)$	$-y$
$\text{CONJ}(y)$	y'

Table III: A collection of reversible instructions, “.” is the broadcasting operations in Julia. The second argument of MULINT and DIVINT should be a nonzero integer.

1. Backward rules for instructions

For function $\vec{y} = f(\vec{x})$, its Jacobian is $J_{ij} = \frac{\partial y_i}{\partial x_j}$ and its Hessian is $H_{ij}^k = \frac{\partial^2 y_k}{\partial x_i \partial x_j}$. We have the following local Jacobians and Hessians on the above instructions.

1. $a += b$

$$J = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$H = \mathbf{0}$$

The inverse is $a -= b$, its Jacobian is the inverse of the matrix above

$$J(f^{-1}) = J^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

In the following, we omit the Jacobians and Hessians of inverse functions.

2. $a += b * c$

$$J = \begin{pmatrix} 1 & c & b \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$H_{bc}^a = H_{cb}^a = 1, \text{ else } 0$$

3. $a += b/c$

$$J = \begin{pmatrix} 1 & 1/c & -b/c^2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$H_{cc}^a = 2b/c^3,$$

$$H_{bc}^a = H_{cb}^a = -1/c^2, \text{ else } 0$$

4. $a += b^c$

$$J = \begin{pmatrix} 1 & cb^{c-1} & b^c \log b \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$H_{bc}^a = H_{cb}^a = b^{c-1} + cb^{c-1} \log b,$$

$$H_{bb}^a = (c-1)cb^{c-2},$$

$$H_{cc}^a = b^c \log^2 b, \text{ else } 0$$

5. $a += e^b$

$$J = \begin{pmatrix} 1 & e^b \\ 0 & 1 \end{pmatrix}$$

$$H_{bb}^a = e^b, \text{ else } 0$$

6. $a += \log b$

$$J = \begin{pmatrix} 1 & 1/b \\ 0 & 1 \end{pmatrix}$$

$$H_{bb}^a = -1/b^2, \text{ else } 0$$

7. $a += \sin b$

$$J = \begin{pmatrix} 1 & \cos b \\ 0 & 1 \end{pmatrix}$$

$$H_{bb}^a = -\sin b, \text{ else } 0$$

8. $a += \cos b$

$$J = \begin{pmatrix} 1 & -\sin b \\ 0 & 1 \end{pmatrix}$$

$$H_{bb}^a = -\cos b, \text{ else } 0$$

9. $a \leftarrow |b|$

$$J = \begin{pmatrix} 1 & \text{sign}(b) \\ 0 & 1 \end{pmatrix}$$

$$H = \mathbf{0}$$

10. $a \leftarrow -a$

$$J = \begin{pmatrix} -1 \end{pmatrix}$$

$$H = \mathbf{0}$$

11. $\text{SWAP}(a, b) = (b, a)$

$$J = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$H = \mathbf{0}$$

12.

$$\text{ROT}(a, b, \theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

$$J = \begin{pmatrix} \cos \theta & -\sin \theta & -b \cos \theta - a \sin \theta \\ \sin \theta & \cos \theta & a \cos \theta - b \sin \theta \\ 0 & 0 & 1 \end{pmatrix}$$

$$H_{a\theta}^a = H_{\theta,a}^a = -\sin \theta,$$

$$H_{b\theta}^a = H_{\theta,b}^a = -\cos \theta,$$

$$H_{\theta\theta}^a = -a \cos \theta + b \sin \theta,$$

$$H_{a\theta}^b = H_{\theta,a}^b = \cos \theta,$$

$$H_{b\theta}^b = H_{\theta,b}^b = -\sin \theta,$$

$$H_{\theta\theta}^b = -b \cos \theta - a \sin \theta, \text{ else } 0$$

Appendix C: Learn by consistency

Consider a training that with input \mathbf{x}^* and output \mathbf{y}^* , find a set of parameters \mathbf{p}_x that satisfy $\mathbf{y}^* = f(\mathbf{x}^*, \mathbf{p}_x)$. In traditional machine learning, we define a loss $\mathcal{L} = \text{dist}(\mathbf{y}^*, f(\mathbf{x}^*, \mathbf{p}_x))$ and minimize it with gradient $\frac{\partial \mathcal{L}}{\partial \mathbf{p}_x}$. This works only when the target function is locally differentiable.

Here we provide an alternative by making use of reversibility. We construct a reversible program $\mathbf{y}, \mathbf{p}_y = f_r(\mathbf{x}, \mathbf{p}_x)$, where \mathbf{p}_x and \mathbf{p}_y are “parameter” spaces on the input side and output

side. The algorithm can be summarized as

Algorithm 2: Learn by consistency

Result: \mathbf{p}_x

Initialize \mathbf{x} to \mathbf{x}^* , parameter space \mathbf{p}_x to random.

if \mathbf{p}_y is null **then**

$\mathbf{x}, \mathbf{p}_x = f_r^{-1}(\mathbf{y}^*)$

else

$\mathbf{y}, \mathbf{p}_y = f_r(\mathbf{x}, \mathbf{p}_x)$

while $\mathbf{y} \neq \mathbf{y}^*$ **do**

$\mathbf{y} = \mathbf{y}^*$

$\mathbf{x}, \mathbf{p}_x = f_r^{-1}(\mathbf{y}, \mathbf{p}_y)$.

$\mathbf{x} = \mathbf{x}^*$

$\mathbf{y}, \mathbf{p}_y = f_r(\mathbf{x}, \mathbf{p}_x)$

Here, $\text{parameter}(\cdot)$ is a function for taking the parameter space. This algorithm utilizes the self-consistency relation

$$\mathbf{p}_x^* = \text{parameter}(f_r^{-1}(\mathbf{y}^*, \text{parameter}(f_r(\mathbf{x}^*, \mathbf{p}_x^*)))), \quad (\text{C1})$$

Similar idea of training by consistency is used in self-consistent meanfield theory [76] in physics. Finding the self-consistent relation is crucial to a self-consistency based training. Here, the reversibility provides a natural self-consistency relation. However, it is not a silver bullet, let's consider the following example

```
@i function f1(y!, x, p!)
    p! += identity(x)
    y! -= exp(x)
    y! += exp(p!)
end

@i function f2(y!, x!, p!)
    p! += identity(x!)
    y! -= exp(x!)
    x! -= log(-y!)
    y! += exp(p!)
end

function train(f)
    loss = Float64[]
    p = 1.6
    for i=1:100
        y!, x = 0.0, 0.3
        @instr f(y!, x, p)
        push!(loss, y!)
        y! = 1.0
        @instr (~f)(y!, x, p)
    end
    loss
end
```

Functions `f1` and `f2` computes $f(x, p) = e^{(p+x)} - e^x$ and stores the output in a new memory `y!`. The only difference is `f2` uncomputes `x` arithmetically. The task of training is to find a `p` that make the output value equal to target value 1. After 100 steps, `f2` runs into the fixed point with `x` equal

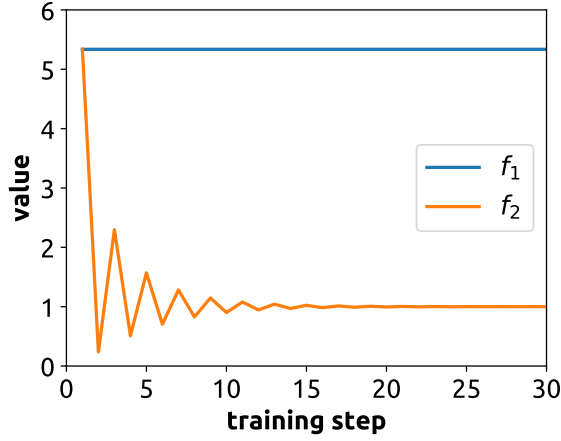


Figure 8: The output value $y!$ as a function of self-consistent training step.

to 1 upto machine precision. However, parameters in f_1 does change at all. The training of f_1 fails because this function actually computes $f_1(y, x, p) = y + e^{(p+x)} - e^x, x, x+p$, where the training parameter p is completely determined by the parameter space on the output side $x \cup x + p$. As a result, shifting y directly is the only approach to satisfy the consistency relation. On the other side, $f_2(y, x, p) = y + e^{(p+x)} - e^x, \tilde{0}, x + p$, the output parameters $\tilde{0} \cup x + p$ can not uniquely determine input parameters p and x . Here, we use $\tilde{0}$ to denote the zero with rounding error.

By viewing \mathbf{x} and parameters in \mathbf{p}_x as variables, we can study the trainability from the information perspective.

Theorem 2. Only if the conditional entropy $S(\mathbf{y}|\mathbf{p}_y)$ is nonzero, algorithm 2 is trainable.

Proof. The above example reveals a fact that training impossible when output parameters completely determines input parameters (or $S(\mathbf{p}_x|\mathbf{p}_y) = 0$).

$$\begin{aligned}
 S(\mathbf{p}_x|\mathbf{p}_y) &= S(\mathbf{p}_x \cup \mathbf{p}_y) - S(\mathbf{p}_y) \\
 &\leq S((\mathbf{p}_x \cup \mathbf{x}) \cup \mathbf{p}_y) - S(\mathbf{p}_y), \\
 &\leq S((\mathbf{p}_y \cup \mathbf{y}) \cup \mathbf{p}_y) - S(\mathbf{p}_y), \\
 &\leq S(\mathbf{y}|\mathbf{p}_y).
 \end{aligned} \tag{C2}$$

The third line uses the bijectivity $S(\mathbf{x} \cup \mathbf{p}_x) = S(\mathbf{y} \cup \mathbf{p}_y)$. This inequality shows that when the parameter space on the output side satisfies $S(\mathbf{y}|\mathbf{p}_y) = 0$, i.e. contains all information to determine the output field, the input parameters are also completely determined by this parameter space, hence training can not work. \square

In the above example, it corresponds to the case $S(e^{(x+y)-e^x} | x \cup x + y) = 0$ in f_1 . The solution is to remove the information redundancy in output parameter space through uncomputing as shown in f_2 .

Appendix D: Functions used in the main text

```

"""
    dot(out!, v1, v2)

dot product.
"""
@i function dot(out!, v1::Vector{T}, v2) where T
    for i = 1:length(v1)
        out! += v1[i]*v2[i]
    end
end

"""
    norm2(out!, vec)

squared norm.
"""
@i function norm2(out!, vec::Vector{T}) where T
    anc1 ← zero(T)
    for i = 1:length(vec)
        anc1 += identity(vec[i]')
        out! += anc1*vec[i]
        anc1 -= identity(vec[i]')
    end
end

```

In `norm2`, we copied `vec[i]'` to `anc1` to avoid the same variable appear twice in the argument list of $\oplus(*)$, where the prime represents the adjoint dataview.

```

"""
    var_and_mean_sq(var!, mean!, sqv)

Variance and mean value from squared values `sqv`.
"""
@i function var_and_mean_sq(var!, mean!, sqv)
    sqmean ← zero(mean!)
    @inbounds for i=1:length(sqv)
        mean! += sqv[i] ^ 0.5
        var! += identity(sqv[i])
    end
    DIVINT(mean!, length(sqv))
    DIVINT(var!, length(sqv))
    sqmean += mean! ^ 2
    var! -= identity(sqmean)
    sqmean -= mean! ^ 2
    MULINT(var!, length(sqv))
    DIVINT(var!, length(sqv)-1)
end

```

Appendix E: CUDA compitibility

CUDA programming is playing a more and more important role in high performance computing. In Julia, one can write kernel functions in native Julia language with CUDA-

native [77]. NiLang is compatible with CUDAnative, one can write a reversible kernel like the following.

```
using CuArrays, CUDAnative, GPUArrays
using NiLang, NiLang.AD

@i @inline function swap_kernel(state, mask1, mask2)
    @invcheckoff b ← (blockIdx().x-1) *
        blockDim().x + threadIdx().x
    @invcheckoff if (b < length(state), ~)
        if (b&mask1==0 && b&mask2==mask2, ~)
            SWAP(state[b+1], state[b ∨
                (mask1|mask2) + 1])
        end
    end
end
```

This kernel function simulates the SWAP gate in quantum

computing. Here, we use the macro `@invcheckoff` to remove the reversibility checks. It is necessary because the possible error thrown in a kernel function can not be handled properly on a CUDA thread. One can launch this kernel function to GPUs with a single macro `@cuda` as show in the following using case.

```
julia> @i function instruct!(state::CuVector,
    gate::Val{:SWAP}, locs::Tuple{Int,Int})
    mask1 ← 1 << (tget(locs, 1)-1)
    mask2 ← 1 << (tget(locs, 2)-1)
    XY ← GPUArrays.thread_blocks_heuristic(
        length(state))
    @cuda threads=tget(XY,1) blocks=tget(XY,
        2) swap_kernel(state, mask1, mask2)
end

julia> instruct!(CuArray(randn(8)),
    Val{:SWAP}, (1,3))[1]
8-element CuArray{Float64,1,Nothing}:
-0.06956048379200473
-0.6464176838567472
-0.06523362834285944
-0.7314356941903547
 1.512329204247244
 0.9773772766637732
 1.6473223915215722
-1.0631789613639087
```