

A Comparison of Automatic Differentiation and Continuous Sensitivity Analysis for Derivatives of Differential Equation Solutions

Christopher Rackauckas^{*1,2,3,4}, Yingbo Ma^{†1,3}, Vaibhav Dixit^{3,5}, Xingjian Guo^{3,6}, Mike Innes⁷, Jarrett Revels^{8,9}, Joakim Nyberg¹⁰, and Vijay Ivaturi³

¹*Department of Mathematics, University of California, Irvine, Irvine, CA 92697, USA*

²*Center for Complex Biological Systems, University of California, Irvine, Irvine, CA 92697, USA*

³*Center for Translational Medicine, University of Maryland, Baltimore, Baltimore, MD 21201, USA*

⁴*Department of Mathematics, Massachusetts Institute of Technology, Boston, MA 02139, USA*

⁵*Department of Mathematical Sciences, Indian Institute of Technology (B.H.U.), Varanasi, Uttar Pradesh 221005 India*

⁶*Courant Institute of Mathematical Sciences, New York University, New York, New York, NY 10012, USA*

⁷*Julia Computing Inc.*

⁸*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

⁹*Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

¹⁰*Department of Pharmaceutical Biosciences, Uppsala University, SE-751 24, Sweden*

Abstract

The derivatives of differential equation solutions are commonly used as model diagnostics and as part of parameter estimation routines. In this manuscript we investigate an implementation of Discrete local Sensitivity Analysis via Automatic Differentiation (DSAAD). A non-stiff Lotka-Volterra model, a discretization of the two dimensional ($N \times N$) Brusselator stiff reaction-diffusion PDE, a stiff non-linear air pollution and a non-stiff pharmacokinetic/pharmacodynamic (PK/PD) model were used as prototype models for this investigation. Our benchmarks show that on sufficiently small (<100 parameters) stiff and non-stiff systems of ODEs, forward-mode DSAAD is more efficient than both reverse-mode DSAAD and continuous forward/adjoint sensitivity analysis. The scalability of continuous adjoint methods is shown to result in better efficiency for larger ODE systems such as PDE discretizations. In addition to testing efficiency, results on test equations demonstrate the applicability of DSAAD to differential-algebraic equations, delay differential equations, and hybrid differential equation systems where the event timing and effects are dependent on model parameters. Together, these results show that language-level automatic differentiation is an efficient method for calculating local sensitivities of a wide range of differential equation models.

1 Introduction

In the literature of differential equations, local sensitivity analysis is the practice of calculating derivatives to a differential equation's solution with respect to model parameters. For an ordinary differential equation (ODE) of the form

$$u' = f(u, p, t), \quad (1)$$

where f is the derivative function and p are the model parameters, the sensitivity of the state vector u with respect to model parameter p_i at time t is defined as $\frac{\partial u(p,t)}{\partial p_i}$. These sensitivities have many applications. For example, they can be directly utilized in fields such as biological modeling to identify parameters of interest for tuning and experimentation [1]. Recent studies have utilized these sensitivities as part for training neural networks associated with the ODEs [2, 3]. In addition, these sensitivities are indirectly utilized in many disciplines for parameter estimation of dynamical models. Parameter estimation is the problem of finding parameters p such that a cost

*crackauc@mit.edu

†Co-first author

function $C(p)$ is minimized (usually some fit against data) [4, 5, 6, 7, 1, 8, 9]. Gradient-based optimization methods require the computation of gradients of $C(p)$. By the chain rule, $\frac{dC}{dp}$ requires the calculation of $\frac{du(t_i)}{dp}$ which are the model sensitivities. Given the high computation cost of parameter estimation due to the number of repeated numerical solutions which are required, efficient and accurate computation of model sensitivities is an important part of differential equation solver software.

The simplest way to calculate model sensitivities is to utilize numerical differentiation which is given by the formula

$$\frac{\partial u(t)}{\partial p_i} = \frac{u(p + \Delta p_i, t) - u(p, t)}{\Delta p_i} + \mathcal{O}(\Delta p_i), \quad (2)$$

where $p + \Delta p_i$ means adding Δp_i to only the i th component of p . However, this method is not efficient (it requires two numerical ODE solutions for each parameter i) and it is prone to numerical error. If Δp_i is chosen too large, then the error term of the approximation is large. In contrast, if Δp_i is chosen too small, then calculations may exhibit floating point cancellation which increases the error [10].

To alleviate these issues, many differential equation solver softwares implement a form of sensitivity calculation called continuous local sensitivity analysis (CSA) [11, 12]. Forward-mode continuous sensitivity analysis calculates the model sensitivities by extending the ODE system to include the equations:

$$\frac{d}{dt} \left(\frac{\partial u}{\partial p_i} \right) = \frac{\partial f}{\partial u} \frac{\partial u}{\partial p_i} + \frac{\partial f}{\partial p_i} \quad (3)$$

where $\frac{\partial f}{\partial u}$ is the Jacobian of the derivative function f with respect to the current state, and $\frac{\partial f}{\partial p_i}$ is the gradient of the derivative function with respect to the i th parameter. Since these equations for each i are dependent on the current state u , these ODEs must be solved simultaneously with the ODE system $u' = f$. By solving this expanded system, one can ensure that the sensitivities are computed to the same error tolerance as the original ODE terms, and only a single numerical ODE solver call is required.

However, since the number of ODEs in this system now scales proportionally with the number of parameters, forward-mode CSA is not practical for a large number of parameters. Instead, for these cases continuous adjoint sensitivity analysis (CASA) is utilized. This methodology is defined to directly compute the gradient of a cost function of the solution. Given a cost function on the ODE solution which is evaluated as discrete time points (such as an L^2 loss)

$$C(u(p)) = \sum_i c(u(p, t_i)) \quad (4)$$

this is done by solving a backwards ODE known as adjoint problem

$$\frac{d\lambda^*}{dt} = -\lambda^* \frac{\partial f(u(t), p, t)}{\partial u} \quad (5)$$

where at every time point t_i , this backwards ODE is perturbed by $\frac{\partial c(u(p, t_i))}{\partial u}$. Note $u(t)$ is generated by a forward solution. The gradient of the cost function is then given by the integral:

$$\frac{dC}{dp} = \lambda^*(t_0) \frac{\partial f(u(t_0), p, t_0)}{\partial p} + \sum_i \int_{t_i}^{t_{i+1}} \lambda^*(t) \frac{\partial f(u(t), p, t)}{\partial p} dt \quad (6)$$

which is evaluated via numerical quadrature.

In contrast to CSA methods, discrete sensitivity analysis calculates model sensitivities by directly differentiating the numerical method's steps [11]. However, this approach requires specialized implementations of the first order ODE solvers to propagate said derivatives.

Instead, one can achieve the same end by using automatic differentiation (AD) on a solver implemented entirely in a language with pervasive AD. Section 2 introduces the discrete sensitivity analysis through AD (DSAAD) approach via type-specialization on a generic algorithm. Section 3 compares the performance of DSAAD against continuous sensitivity analysis and numerical differentiation approaches and shows that DSAAD consistently performs well on the tested models. Section 5 describes limitations of the continuous sensitivity analysis approach and describes how these cases are automatically handled in the case of DSAAD. Together, this manuscript shows that the ability to utilize AD directly on numerical integrator can be advantageous to existing approaches for the calculation of model sensitivities.

2 Discrete Sensitivity Analysis Through Automatic Differentiation

The core feature of the Julia programming language is multiple dispatch [13]. It allows a function to compile to different outputs dependent on the types of the inputs, effectively allowing choices of input types to trigger forms of code generation. ForwardDiff.jl provides a Dual number type which performs automatic differentiation on differentiable programs by simultaneously propagating a derivative along with the computed value on atomic (addition, multiplication, etc.) and standard mathematical (sin, exp, etc.) function calls [14]. By utilizing the chain rule during the propagation, any function which is composed of differentiable calls is also differentiable by the methodology. Such a program is known as a differentiable program. Since this exactly differentiates the atomics, the numerical error associated with this method is similar to the standard evaluation of the function, effectively alleviating the errors seen in numerical differentiation. In addition, this method calculates derivatives simultaneously with the function’s evaluation, making it a good candidate for fast discrete sensitivity analysis. It can be thought of as the analogue to continuous forward-mode sensitivity analysis on the set of differentiable programs. Similarly, reverse-mode AD from packages like ReverseDiff.jl [15] and Flux.jl [16]. utilize Tracker numerical types which builds a tape of the operations and utilizes the reverse of the chain rule to “backpropagate” derivatives and directly calculate the gradient of some cost function. This implementation of AD can be thought of as the analogue of CASA on the set of differentiable programs.

The DifferentialEquations.jl package provides many integration routines which were developed in native Julia [17]. These methods are type-generic, meaning they utilize the numeric and array types that are supplied by the user. Thus these ODE solvers serve as a generic template whose internal operations can be modified by external packages via dispatch. When a generic DifferentialEquations.jl integrator is called with a Dual number type for the initial condition, the combination of these two programs results in a program which performs discrete sensitivity analysis. This combination is what we define as forward-mode DSAAD, and this combination with reverse-mode ReverseDiff.jl AD Tracker types is reverse-mode DSAAD.

To test the correctness of the DSAAD method, we check the outputted sensitivities on four models. For our tests we utilize nonlinear reaction models which are representative of those found in biological and pharmacological applications where these techniques are commonly used [18, 19]. The models are:

1. The non-stiff Lotka-Volterra equations (LV).
2. An $N \times N$ finite difference discretization of the two-dimensional stiff Brusselator reaction-diffusion PDE (BRUSS).
3. A stiff pollution model (POLLU).
4. A non-stiff pharmacokinetic/pharmacodynamic system (PK/PD).

The cover stiff and non-stiff ODEs, large systems and small systems, and include the a PDE discretization with a dimension N for testing the scaling of the methodologies. Details of the four models are presented in the Appendix.

Figure 1 shows the output of the first two models’ sensitivities that are computed by the DSAAD method compared to CSA. The two methods align in their model sensitivity calculations, demonstrating that the application of AD on the generic ODE solver does produce correct output sensitivities. From these tests we note that the differences in the model sensitivities between the two methods had a maximum norm of $1.14 \cdot 10^{-5}$ and $3.10 \cdot 10^{-4}$ which is roughly the chosen tolerance of the numerical integration.

3 Discrete and Continuous Sensitivity Analysis Performance Comparisons

3.1 Forward-Mode Sensitivity

To test the relative performance of discrete and continuous sensitivity analysis, we utilized packages from the Julia programming language. The method for continuous sensitivity analysis is implemented in the DiffEqSensitivity.jl package by directly extending a user-given ordinary differential equation. This is done by extending the initial

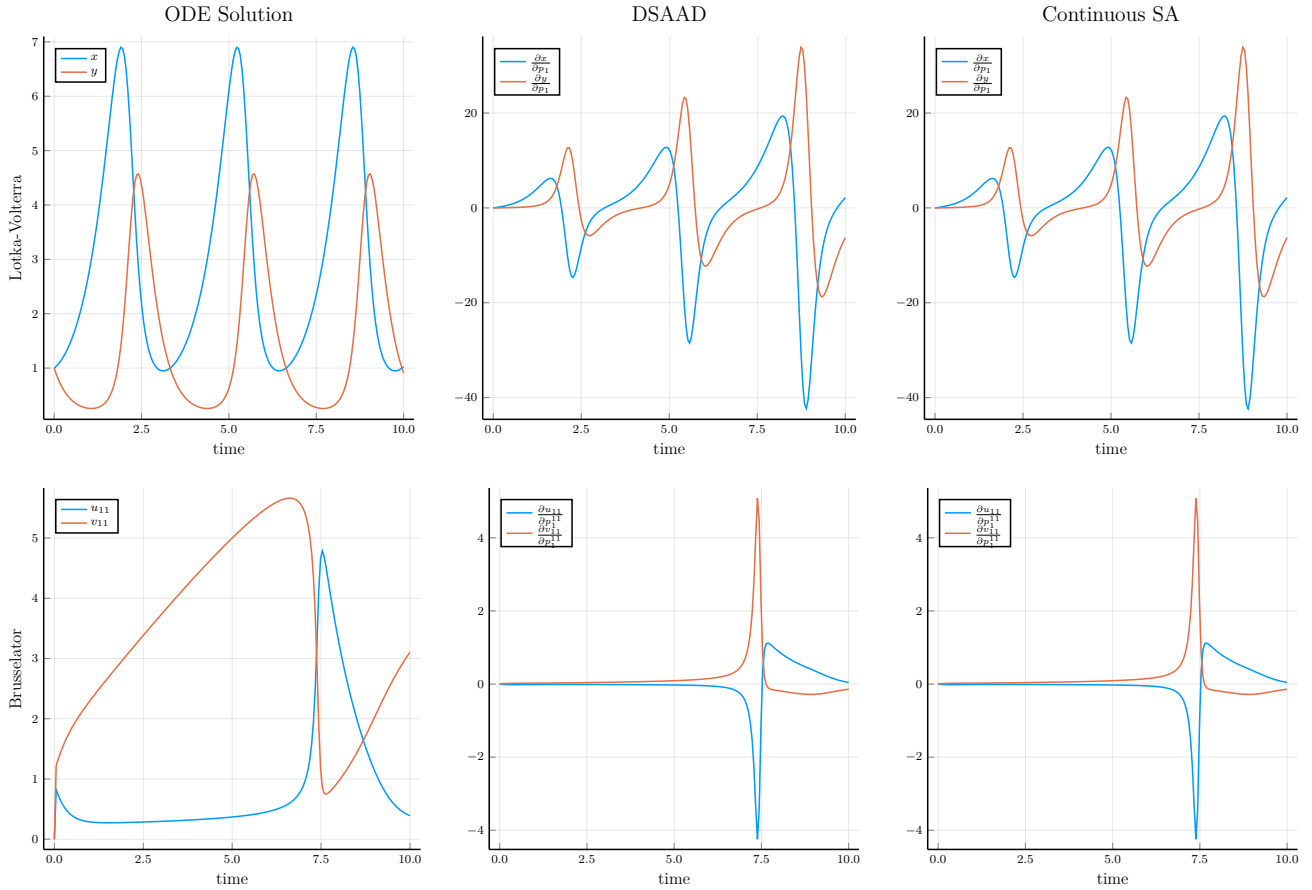


Figure 1: **Model Sensitivities for DSAAD and CSA.** **Top row:** For the Lotka-Volterra model, the results are shown for $t \in [0, 10]$. **Bottom row:** The Brusselator PDE is discretized using a 3×3 uniform grid on the domain $[0, 1] \times [0, 1]$, and the resulting ODE system is solved for $t \in [0, 10]$. The results for u and v on the $(1, 1)$ grid point are shown.

Method/Runtime	LV (μs)	BRUSS (s)	POLLU (s)	PKPD (ms)
DSAAD	120	0.65	0.015	1.64
CSA User-Jacobian	180	140	0.71	4.75
CSA AD-Jacobian	290	150	0.81	5.20
CSA AD- Jv seeding	230	160	0.71	5.66
Numerical Differentiation	670	3.9	0.036	12.3

Table 1: **Forward Sensitivity Analysis Performance Benchmarks.** The Lotka-Volterra model used by the benchmarks is the same as in Figure 1, while the Brusselator benchmarks use a finer 5×5 grid with the same solution domain, initial values and parameters. The `Tsit5` integrator is used for the Lotka-Volterra and the PKPD model. The `Rodas5` integrator is used for the Brusselator and POLLU.

Method/Runtime	LV (ms)	BRUSS (s)	POLLU (s)	PKPD (ms)
Forward-Mode DSAAD	0.23	0.66	0.01	5.21
Reverse-Mode DSAAD	3.1	11.9	0.91	195
CASA User-Jacobian	6.9	0.48	2.31	34.5
CASA AD-Jacobian	7.1	0.96	3.24	34.4
CASA AD- $v'J$ seeding	7.7	1.81	7.33	49.2
Numerical Differentiation	1.1	4.19	0.05	31.2

Table 2: **Adjoint Sensitivity Analysis Performance Benchmarks.** The Lotka-Volterra and Brusselator models used by the benchmarks are the same as in Figure 1. The integrators used for the benchmarks are: `Rodas5` for Brusselator and POLLU, and `Tsit5` for Lotka-Volterra and PKPD.

condition vector and defining a new derivative function \tilde{f} which performs f on the first N components and adds the sensitivity equations. For performance, construction of the Jacobian can be avoided by utilizing AD for vector-Jacobian and Jacobian-vector products. By seeding the Dual numbers to have partials v and applying forward applications of f , the resulting output is the desired $\frac{\partial f}{\partial u} v$. Similarly, seeding on Tracked reals for reverse-mode autodifferentiation results in $v \frac{\partial f}{\partial u}^T$ which is the other desired quantity. As a comparison, full Jacobian implementations were also explored, either via a user-given analytical solution or automatic differentiation. As a comparison, we also include numerical differentiation performed by the `DiffEqDiffTools.jl` package.

The test problems were solved with the various sensitivity analysis methods and the timings are given in Table 1. In all of these benchmarks DSAAD performs well, being the fastest in all cases. We note that in all of these cases a dense Jacobian was used for the CSA, which in cases with moderately sized (20x20) sparse Jacobians led to performance disadvantages for CSA while DSAAD naturally builds a sparse representation of the differentiation by directly interleaving the derivative calculations with the original operations. For the BRUSS and Pollution problems, this Jacobian vector product was performed at runtime which is one of the reasons for its decreased efficiency.

3.2 Adjoint Sensitivity

For adjoint sensitivity analysis, discrete adjoint sensitivity analysis programs were produced using a combination of the generic `DifferentialEquations.jl` integrator with the tape-based automatic differentiation implementation of `ReverseDiff.jl`. `DiffEqSensitivity.jl` provides an implementation of CASA which saves a continuous solution for the forward pass of the solution and utilizes its interpolant in order to calculate the requisite Jacobian and gradients for the backwards pass. While this method is less memory efficient than checkpointing or re-solving schemes [12], it only requires a single forward numerical solution and thus is more runtime optimized for sufficiently small models.

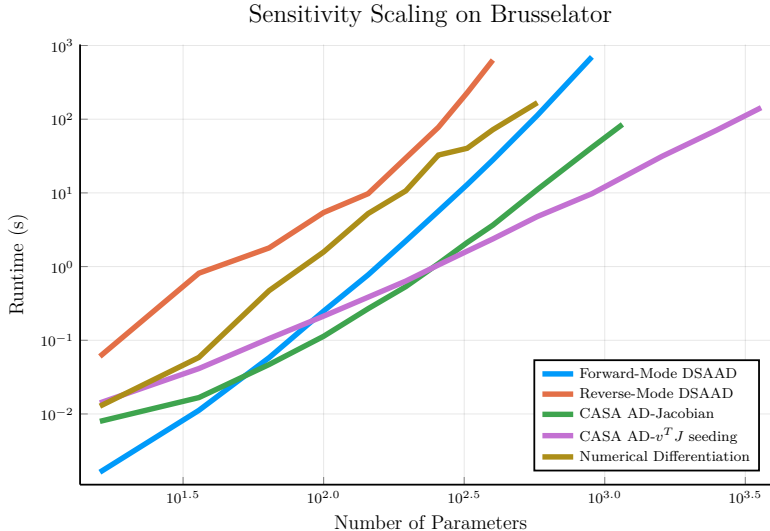


Figure 2: **Brusselator Scaling Benchmarks.** The Brusselator problem was solved with varying dimension N with each of the different sensitivity analysis methods to determine their scalability with respect to number of parameters. Depicted on the x-axis is the number of parameters $4N^2$ in log-scale with the y-axis being the runtime in seconds in log-scale.

The timing results of these methods on the test problems are given in Table 2. These results show a clear performance advantage for forward-mode DSAAD over the other choices on sufficiently small models.

4 Adjoint Sensitivity Scaling

The previous tests all showed that on small models forward-mode via AD was advantageous to reverse-mode and adjoint methods. However, the advantage of adjoint methods comes in their ability to scale, scaling with the number of terms in the differential equation as opposed to the number of parameters. Thus we decided to test the scaling of the methods on the Brusselator problem. For an $N \times N$ discretization in space, this problem has $2N^2$ ODE terms and $4N^2$ parameters. The timing results for the adjoint methods and forward-mode DSAAD are shown in Figure 2. This demonstrates that as N increases, there is a point at which CASA becomes more efficient than DSAAD. We note that the inability for the reverse-mode DSAAD to scale comes from current implementation issues. The Tracker types in the differential equation solvers have to utilize scalar tracking instead of tracking array primitives, greatly increasing the size of the computational graph and the memory burden. When array primitives are used, mutation is not allowed which decreases the efficiency of the solver more than the gained efficiency. These issues may be addressed in the next generation reverse-mode source-to-source AD packages like Zygote [20] or Capstan [21] by not relying on type propagation.

4.1 Application: Parameter Estimation

To test the effect of sensitivity analysis timings in applications, we benchmarked parameter estimation performed with an L^2 loss function on generated data for each of the models. The data was generated by solving the ODE with the parameters defined in the Appendix and sampling at evenly spaced time points (100 points for Lotka-Volterra, 20 for Brusselator, 10 for POLLU, and 41 for PK/PD). Each of the parameter estimations were done using the BFGS local optimizer from Optim.jl [22] and were ran until the optimizer converged to the optima to a tolerance of 10^{-6} (results were checked for proper convergence). Each method started from the same initial condition which was a perturbation of the true parameters. For Lotka-Volterra, the initial parameter values were $\frac{4}{5}$ the true values, for PK/PD $0.95p_0 + 0.001$, and for the other models $\frac{9}{10}$. The timings are shown in

Method/Runtime	LV (<i>s</i>)	BRUSS (<i>s</i>)	POLLU (<i>s</i>)	PKPD (<i>s</i>)
Forward-Mode DSAAD	0.125	0.617	0.065	2.07
CSA User-Jacobian	0.162	47.7	4.05	3.77
CSA AD-Jacobian	0.152	56.5	3.92	6.27
CSA AD- <i>Jv</i> seeding	0.159	56.3	3.69	6.25
Reverse-Mode DSAAD	0.223	44.7	4.59	85.0
CASA User-Jacobian	0.425	0.708	6.23	60.3
CASA AD-Jacobian	0.418	0.605	3.08	59.8
CASA AD- <i>v'J</i> seeding	0.436	1.52	7.78	61.1
Numerical Differentiation	0.096	4.71	0.281	8.53

Table 3: **Parameter Estimation Benchmarks.** The Lotka-Volterra model used by the benchmarks is the same as in Figure 1, while the Brusselator benchmarks use a finer 5×5 grid with the same solution domain, initial values and parameters. The `Tsit5` integrator is used for Lotka-Volterra and PKPD, and the `Rodas5` integrator is used for Brusselator and pollution.

Table 3. While not as pronounced as the pure sensitivity calculations, these benchmarks show that utilizing a more efficient sensitivity analysis calculation does give a performance advantage in the application. Additional performance disadvantages for numerical differentiation could be attributed to the increased numerical error in the gradient which notably caused more iterations for the optimizer.

5 Discrete Sensitivity Analysis Through Automatic Differentiation Readily Generalizes To Hybrid, Delay, and Differential-Algebraic Differential Equations

We compared the flexibility of sensitivity analysis approaches in order to understand their relative merits for use in a general-purpose differential equation package. First we analyze the ability for the sensitivity analysis approaches to work with event handling. Event-handling is a feature of differential equation packages which allows users to provide a rootfinding function $g(u, p, t)$ at which a discontinuity (of the user’s choice) is applied at every time point where $g(u, p, t) = 0$ (such equations are also known as hybrid differential equations). Automatic differentiation approaches to discrete sensitivity analysis directly generalize to handling this case by propagating Dual numbers through the event handling code. On the other hand, continuous sensitivity analysis approaches can require special handling in order to achieve correctness. There are two ways which sensitivities will not propagate:

1. Standard continuous sensitivity analysis does not take into account the sensitivity of the time point of the discontinuity to the parameters.
2. Standard continuous sensitivity analysis does not take into account the possibility of the discontinuity’s amount being parameter-dependent.

These points can be illustrated using a single state linear control problem where x is the signal responsible for the control of y . This results in a first-order linear hybrid ordinary differential equation system:

$$\begin{aligned} \frac{dx}{dt} &= -a, \\ \frac{dy}{dt} &= b, \end{aligned} \tag{7}$$

where $a, b > 0$ are parameters, and the rootfinding function is $g(x, y, p, t) = x$. At zero-crossings the parameter b is set to 0, effectively turning off the second equation. For the initial condition $(x(0), y(0)) = (1, 0)$, the system’s

Method	$\partial x(1)/\partial a$	$\partial y(1)/\partial a$	$\partial x(1)/\partial b$	$\partial y(1)/\partial b$
Analytical Solution	-1.0	-0.25	0	0.5
DSAAD	-1.0	-0.25	5.50e-11	0.5
CSA	-1.0	0.0	0.0	1.0

Table 4: **Sensitivity Analysis with Events.** Shown are the results of the control problem given by Equation 7 with $a = 2$ and $b = 1$. It was solved on $t \in [0, 1]$ with initial condition $(x(0), y(0)) = (1, 0)$. The sensitivities of the two state variables are given at time $t = 1$ with respect to the two parameters. The analytical solution is derived by taking derivatives directly on Equation 8.

analytical solution is

$$\begin{aligned}
 x(t) &= 1 - at, \\
 y(t) &= \begin{cases} bt & (t < t^*) \\ bt^* = \frac{b}{a} & (t \geq t^*), \end{cases} \quad (8)
 \end{aligned}$$

where $t^* = 1/a$ is the crossing time, which depends on the parameters. Furthermore, the amount of jump for the discontinuity of dy/dt is also parameter dependent.

The sensitivity analysis results are compared to the true derivative at time $t = 1$ utilizing the analytical solution of the system (Equation 8) in Table 4. These results show that continuous sensitivity analysis as defined in Equation 3 does not properly propagate the sensitivities due to discontinuities and this results in incorrect derivative calculations for hybrid ODE systems. Extensions to continuous sensitivity analysis are required in order to correct these errors which are have been previously detailed [23].

Additionally, the continuous sensitivity analysis equations defined in Equation 3 only apply to ordinary differential equations. It has been shown that a different set of equations is required for delay differential equations (DDEs) [24]

$$\frac{d}{dt} \frac{\partial u(t)}{\partial p_i} = \frac{\partial G}{\partial u} \frac{\partial u}{\partial p_i}(t) + \frac{\partial G}{\partial \tilde{u}} \frac{\partial u}{\partial p_i}(t - \tau) + \frac{\partial G}{\partial p_i}(t), \quad (9)$$

where $\frac{du(t)}{dt} = G(u, \tilde{u}, p, t)$ is the DDE system with a single fixed time delay τ , and differential-algebraic equations (DAEs) [12]

$$\frac{\partial F}{\partial u} \frac{\partial u}{\partial p_i} + \frac{\partial F}{\partial u'} \frac{\partial u'}{\partial p_i} + \frac{\partial F}{\partial p_i} = 0, \quad (10)$$

where $F(u', u, p, t) = 0$ is the DAE system. On the other hand, the discrete sensitivity analysis approach implemented via automatic differentiation is not specialized to ordinary differentiation equations since it automatically generates the sensitivity propagation at the compiler-level utilizing the atomic operations inside the numerical integration scheme. Thus these types of equations, and expanded forms such as DDEs with state-dependent delays or hybrid DAEs, are automatically supported by the connection between DifferentialEquations.jl and Julia-based automatic differentiation packages. Tests on the DDE and DAE solvers confirm this to be the case.

6 Discussion

Performant and correct sensitivity analysis is crucial to many applications of differential equation models. Here we analyzed both the performance and generalizability of the two approaches. Our results show a strong performance advantage for automatic differentiation based discrete sensitivity analysis for forward-mode sensitivity analysis, and an advantage in adjoint sensitivity analysis for sufficiently small systems. Performance profiling shows that the limiting factor to the continuous sensitivity analysis approach is the Jacobian-vector calculation in the sensitivity ODEs, while AD approaches can remove this issue entirely. Notably, reverse-mode automatic differentiation did not perform or scale well in these benchmarks. The implementations have been generally optimized for the usage in machine learning models and more work will need to be done for them to show performance advantages in the context of a solver for stiff ODEs.

One major result to note is that, in many cases of interest, runtime overhead of more rigorous methods can be larger than the theoretical scaling advantages. Forward-mode automatic differentiation does not exhibit the best scaling properties but on ODEs with small numbers of parameters, both stiff and non-stiff, this method benchmarks as advantageous. On problems like PDEs, the scaling of adjoint methods do confer an advantage to them when the problem is sufficiently large but also a disadvantage when the problem is small. Additionally, the ability to seed automatic differentiation for the Jacobian vector multiplications only was demonstrated as advantageous on the larger problems. We note that we saw a major improvement to the seeded Jacobian vector multiplications by precompiling the type and not utilizing the tracing on the derivative function. This shows that reverse-mode automatic differentiation with static computational graphs can be advantageous over purely dynamic implementations in this context. Together, these results show that having the ability to choose between these different methods is essential for a software wishing to support these separate use cases.

While runtime performance of these methods is usually of interest, it is important to note the memory scaling for the various adjoint sensitivity analysis implementations. The chosen implementation for continuous adjoint sensitivity analysis utilizes a continuous solution to the original ODE in order to generate the adjoint Jacobian and gradients on-demand. This setup only requires a single forward ODE solve but makes a tradeoff due to the high memory requirement to save the full timeseries solution and its interpolating function. For example, with the chosen 9th order explicit Runge-Kutta method due to Verner, the total memory cost is $26NM$ since 26 internal derivative calculations of size N are utilized to construct the interpolant where M is the number of time points. In contrast, checkpointing-based adjoint sensitivity analysis implementations [12] re-solve the forward ODE from a saved time point (a checkpoint) in order to get the u value required for the Jacobian and gradient calculation, increasing the runtime cost while decreasing the memory cost to NC where C is the number of checkpoints. For large PDE applications, the memory usage can be a limiting factor and thus the checkpointing-based implementation can be preferred. For the AD approaches, forward-mode AD for adjoint sensitivities is memory conservative since values along the timeseries are not required to be saved. Instead, the total memory cost is greater by then the result saveless solve by a factor P since each parameter requires a unique partial for the Dual numbers to propagate. In contrast, reverse-mode AD approaches require constructing a tape for the entire initial forward solution and thus has memory scaling similar to the interpolant-based continuous adjoint sensitivity analysis. Checkpointing systems for reverse-mode AD could alleviate these memory issues.

In addition, the continuous solution of λ was stored to perform the quadrature for $\frac{dC}{dp}$. Other implementations can reduce the memory cost here by appending the adjoint ODE with this integral calculation to simultaneously compute the gradients. However, our tests indicated that this method is less computationally efficient (around 2x slower on the POLLU model) because the post-adjoint quadrature was able to significantly reduce the number of $\frac{dC}{dp}$ calculations via the adaptive Gauss-Kronrod quadrature implementation of QuadGK.jl. Thus this once again highlights the tradeoff between runtime and memory efficiency.

These results show many advantages for the AD-based discrete sensitivity analysis. However, there are significant engineering challenges to the development of such integration schemes. In order for this methodology to exist, a general function automatic differentiation tool must exist for the programming language and the entire ODE solver must be compatible with the AD software. This works for the Julia-based DifferentialEquations.jl software since it contains a large number of high performance native-Julia solver implementations. However, many other solver ecosystems do not have such possibilities. For example, common open source packages for solving ordinary differential equation systems include deSolve in R [25] and SciPy for Python [26]. While general automatic differentiation tools exist for these languages ([27], [28]), both of these package call out to Fortran-based ODE solvers such as LSODA [29], and thus AD cannot be directly applied to the solver calls.

Lastly, when considering the maintenance of large software ecosystems the AD-based discrete sensitivity analysis approach gives many advantages. For one, almost no additional code was required to be written by the differential equation community in order for this implementation to exist since it directly works via code generation at compile-time on the generic functions of DifferentialEquations.jl. But an additional advantage is that this same technique applies to the native hybrid, delay, and differential-algebraic integrator present in the library. DifferentialEquations.jl also allows for many other actions to occur in the events. For example, the user can change the number of ODEs during an event, and events can change solver internals like the current integration time. Continuous sensitivity analysis would require a separate implementation for each of these equations with could be costly to developer time.

Lastly, we note that this method is composable among pure-Julia codes, meaning that if AD is applied to

a code that calls the numerical differential equation solver, the differentiation of the solver will automatically be computed via the DSAAD method. This means that domain-specific software built using the pure-Julia differential equation solvers would automatically have access to efficient likelihood gradient calculations without having to write or maintain extra code. We are developing software for Pharmaceutical Modeling And Simulation (Pumas.jl) add reference here that performs pharmacometric simulation and non-linear mixed effects (NLME) estimation. The PK/PD model used in this research was built in PuMaS.jl. Our results show that building software such as PuMaS.jl in Julia and utilizing the Julia-based ODE solvers with direct application of AD would give a system that is flexible with respect to the possible equation types without compromising on efficiency.

7 Acknowledgements

This work is supported by Center for Translational Medicine, University of Maryland Baltimore School of Pharmacy. CR is partially supported by the NSF grant DMS1763272 and a grant from the Simons Foundation (594598, QN).

8 Appendix

8.1 Models

The first test problem is LV, the non-stiff Lotka-Volterra model

$$\begin{aligned}\frac{dx}{dt} &= p_1x - p_2xy, \\ \frac{dy}{dt} &= -p_3y + xy.\end{aligned}\tag{11}$$

with initial condition $[1.0, 1.0]$ and $p = [1.5, 1.0, 3.0]$ [30]. The second model, BRUSS, is the two dimensional ($N \times N$) Brusselator stiff reaction-diffusion PDE:

$$\begin{aligned}\frac{\partial u}{\partial t} &= p_2 + u^2v - (p_1 + 1)u + p_3\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + f(x, y, t), \\ \frac{\partial v}{\partial t} &= p_1u - u^2v + p_4\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right),\end{aligned}\tag{12}$$

where

$$f(x, y, t) = \begin{cases} 5 & \text{if } (x - 0.3)^2 + (y - 0.6)^2 \leq 0.1^2 \text{ and } t \geq 1.1 \\ 0 & \text{else,} \end{cases}\tag{13}$$

with no-flux boundary conditions and $u(0, x, y) = 22(y(1 - y))^{3/2}$ with $v(0, x, y) = 27(x(1 - x))^{3/2}$ [31]. This PDE is discretized to a set of $N \times N \times 2$ ODEs using the finite difference method. The parameters are spatially-dependent, $p_i = p_i(x, y)$, making each discretized p_i a $N \times N$ set of values at each discretization point, giving a total of $4N^2$ parameters. The initial parameter values were the uniform $p_i(x, y) = [3.4, 1.0, 10.0, 10.0]$

The third model, POLLU, simulates air pollution. It is a stiff non-linear ODE system which consists 20 ODEs:

$$\begin{aligned}\frac{du_1}{dt} &= -p_1u_1 - p_{10}u_{11}u_1 - p_{14}u_1u_6 - p_{23}u_1u_4 - \\ &\quad p_{24}u_{19}u_1 + p_2u_2u_4 + p_3u_5u_2 + p_9u_{11}u_2 + \\ &\quad p_{11}u_{13} + p_{12}u_{10}u_2 + p_{22}u_{19} + p_{25}u_{20} \\ \frac{du_2}{dt} &= -p_2u_2u_4 - p_3u_5u_2 - p_9u_{11}u_2 - p_{12}u_{10}u_2 + p_1u_1 + p_{21}u_{19} \\ \frac{du_3}{dt} &= -p_{15}u_3 + p_1u_1 + p_{17}u_4 + p_{19}u_{16} + p_{22}u_{19}\end{aligned}$$

$$\begin{aligned}
\frac{du_4}{dt} &= -p_2u_2u_4 - p_{16}u_4 - p_{17}u_4 - p_{23}u_1u_4 + p_{15}u_3 \\
\frac{du_5}{dt} &= -p_3u_5u_2 + p_4u_7 + p_4u_7 + p_6u_7u_6 + p_7u_9 + p_{13}u_{14} + p_{20}u_{17}u_6 \\
\frac{du_6}{dt} &= -p_6u_7u_6 - p_8u_9u_6 - p_{14}u_1u_6 - p_{20}u_{17}u_6 + p_3u_5u_2 + p_{18}u_{16} + p_{18}u_{16} \\
\frac{du_7}{dt} &= -p_4u_7 - p_5u_7 - p_6u_7u_6 + p_{13}u_{14} \\
\frac{du_8}{dt} &= p_4u_7 + p_5u_7 + p_6u_7u_6 + p_7u_9 \\
\frac{du_9}{dt} &= -p_7u_9 - p_8u_9u_6 \\
\frac{du_{10}}{dt} &= -p_{12}u_{10}u_2 + p_7u_9 + p_9u_{11}u_2 \\
\frac{du_{11}}{dt} &= -p_9u_{11}u_2 - p_{10}u_{11}u_1 + p_8u_9u_6 + p_{11}u_{13} \\
\frac{du_{12}}{dt} &= p_9u_{11}u_2 \\
\frac{du_{13}}{dt} &= -p_{11}u_{13} + p_{10}u_{11}u_1 \\
\frac{du_{14}}{dt} &= -p_{13}u_{14} + p_{12}u_{10}u_2 \\
\frac{du_{15}}{dt} &= p_{14}u_1u_6 \\
\frac{du_{16}}{dt} &= -p_{18}u_{16} - p_{19}u_{16} + p_{16}u_4 \\
\frac{du_{17}}{dt} &= -p_{20}u_{17}u_6 \\
\frac{du_{18}}{dt} &= p_{20}u_{17}u_6 \\
\frac{du_{19}}{dt} &= -p_{21}u_{19} - p_{22}u_{19} - p_{24}u_{19}u_1 + p_{23}u_1u_4 + p_{25}u_{20} \\
\frac{du_{20}}{dt} &= -p_{25}u_{20} + p_{24}u_{19}u_1
\end{aligned} \tag{14}$$

with the initial condition of $u_0 = [0, 0.2, 0, 0.04, 0, 0, 0.1, 0.3, 0.01, 0, 0, 0, 0, 0, 0.007, 0, 0, 0]^T$ and parameters $[\cdot 35, 26.6, 12, 300, \cdot 00086, \cdot 00082, 15, 000, \cdot 00013, 24, 000, 16, 500, 9, 000, \cdot 022, 12, 000, 1.88, 16, 300, 4, 800, 000, \cdot 00035, \cdot 0175, 10^9, \cdot 444 \times 10^{12}, 1, 240, 2.1, 5.78, \cdot 0474, 1, 780, 3.12]$ [31].

The fourth model is a non-stiff pharmacokinetic/pharmacodynamic model (PKPD) [32], which is in the form

of

$$\begin{aligned}
\frac{dDepot}{dt} &= -k_a Depot \\
\frac{dCent}{dt} &= k_a Depot + \\
&\quad (CL + V_{max}/(K_m + (Cent/V_c)) + Q_1)(Cent/V_c) + \\
&\quad Q_1(Periph_1/V_{p1}) - \\
&\quad Q_2(Cent/V_c) + Q_2(Periph_2/V_{p2}) \\
\frac{dPeriph_1}{dt} &= Q_1(Cent/V_c) - Q_1(Periph_1/V_{p1}) \\
\frac{dPeriph_2}{dt} &= Q_2(Cent/V_c) - Q_2(Periph_2/V_{p2}) \\
\frac{dResp}{dt} &= k_{in}(1 - (I_{max}(Cent/V_c)^\gamma / (IC_{50}^\gamma + (Cent/V_c)^\gamma))) - k_{out} Resp.
\end{aligned} \tag{15}$$

with the initial condition of [100.0, 0.0, 0.0, 0.0, 5.0]. $k_a = 1$ is the absorption rate of drug into the central compartment from the dosing compartment, $CL = 1$ is the clearance parameter of drug elimination, $V_c = 20$ is the central volume of distribution, $Q_1 = 2$ is the inter-compartmental clearance between central and first peripheral compartment, $Q_2 = 0.5$ is the inter-compartmental clearance between central and second peripheral compartment, $V_{p1} = 10$ is the first peripheral compartment distribution volume, $V_{p2} = 100$ is the second peripheral compartment distribution volume, $V_{max} = 0$ is the maximal rate of saturable elimination of drug, $K_m = 2$ is the Michaelis-Mentens constant, $k_{in} = 10$ is the input rate to the response (PD) compartment with a maximal inhibitory effect of $I_{max} = 1$, $IC_{50} = 2$ is a parameter for the concentration at 50% of the effect and $k_{out} = 2$ is the elimination rate of the response, and $\gamma = 1$ is the model sigmoidicity. Additional doses of 100.0 are applied to the *Depot* variable at every 24 time units.

References

- [1] Andreas Sommer. Numerical Methods for Parameter Estimation in Dynamical Systems with Noise. page 282.
- [2] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. *arXiv:1806.07366 [cs, stat]*, June 2018. arXiv: 1806.07366.
- [3] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models. *arXiv:1810.01367 [cs, stat]*, October 2018. arXiv: 1810.01367.
- [4] M. Peifer and J. Timmer. Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting. *IET Systems Biology*, 1:78–88, 2007.
- [5] Franz Hamilton. Parameter Estimation in Differential Equations: A Numerical Study of Shooting Methods. *SIAM Undergraduate Research Online*, 4:16–31, 2011.
- [6] Xiao Zhen. *Parameter Estimation in Differential Equation Based Models*. PhD thesis.
- [7] Christoph Zimmer. Parameter Estimation for Stochastic Models of Biochemical Reactions. *Journal of Computer Science & Systems Biology*, 06, 2013.
- [8] L. E. Friberg, A. Henningsson, H. Maas, L. Nguyen, and M. O. Karlsson. Model of chemotherapy-induced myelosuppression with parameter consistency across drugs. *J Clin Oncol*, 20:4713–21, December 2002.
- [9] B. Steiert, A. Raue, J. Timmer, and C. Kreutz. Experimental design for parameter estimation of gene regulatory networks. *PLoS ONE*, 7:e40052, 2012.

- [10] Richard L. Burden and J. Douglas Faires. *Numerical analysis*. Brooks/Cole, Cengage Learning, Boston, MA, 9th edition, 2011.
- [11] Hong Zhang and Adrian Sandu. FATODE: A Library for Forward, Adjoint, and Tangent Linear Integration of ODEs. *SIAM Journal on Scientific Computing*, 36(5):C504–C523, January 2014.
- [12] Alan C. Hindmarsh, Peter N. Brown, Keith E. Grant, Steven L. Lee, Radu Serban, Dan E. Shumaker, and Carol S. Woodward. SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. *ACM Trans. Math. Softw.*, 31:363–396, 2005.
- [13] J. Bezanson, A. Edelman, S. Karpinski, and V. Shah. Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59:65–98, January 2017.
- [14] Jarrett Revels, Miles Lubin, and Theodore Papamarkou. Forward-Mode Automatic Differentiation in Julia. *arXiv:1607.07892 [cs]*, July 2016. arXiv: 1607.07892.
- [15] Jarrett Revels. ReverseDiff.jl.
- [16] Mike Innes. Flux: Elegant Machine Learning with Julia. *Journal of Open Source Software*, 2018.
- [17] Christopher Rackauckas and Qing Nie. DifferentialEquations.jl - A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia. *Journal of Open Research Software*, 5:15, 2017.
- [18] Hiroaki Kitano. Computational systems biology.
- [19] Meindert Danhof, Elizabeth C. M. de Lange, Oscar E. Della Pasqua, Bart A. Ploeger, and Rob A. Voskuyl. Mechanism-based pharmacokinetic-pharmacodynamic (PK-PD) modeling in translational drug research. *Trends in Pharmacological Sciences*, 29(4):186–191, April 2008.
- [20] Michael Innes. Don't Unroll Adjoint: Differentiating SSA-Form Programs. *eprint arXiv:1810.07951*, page arXiv:1810.07951, October 2018.
- [21] Jarrett Revels. Capstan.jl.
- [22] Patrick Kofod Mogensen and Asbjorn Nilsen Riseth. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software*, 3(24):615, 2018.
- [23] Christian Kirches. *A Numerical Method for Nonlinear Robust Optimal Control with Implicit Discontinuities and an Application to Powertrain Oscillations*. PhD thesis, October 2006.
- [24] H. T. Banks, Danielle Robbins, and Karyn L. Sutton. Generalized Sensitivity Analysis for Delay Differential Equations. In Kristian Bredies, Christian Clason, Karl Kunisch, and Gregory von Winckel, editors, *Control and Optimization with PDE Constraints*, pages 19–44. Springer Basel, Basel, 2013.
- [25] Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer. Solving Differential Equations in R: Package deSolve. *Journal of Statistical Software*, 33(9), 2010.
- [26] Eric Jones, Travis Oliphant, Pearu Peterson, and others. *SciPy: Open source scientific tools for Python*. 2001.
- [27] Steven E Pav. Madness: a package for Multivariate Automatic Differentiation. page 15.
- [28] Dougal Maclaurin, David Duvenaud, and Ryan P Adams. Autograd: Effortless Gradients in Numpy. page 3.
- [29] Linda Petzold. Automatic Selection of Methods for Solving Stiff and Nonstiff Systems of Ordinary Differential Equations.
- [30] J. D. Murray. *Mathematical biology*. Interdisciplinary applied mathematics. Springer, New York, 3rd ed edition, 2002.

- [31] Ernst Hairer and Gerhard Wanner. *Solving Ordinary Differential Equations II - Stiff and Differential-Algebraic Problems*. Springer, 1991.
- [32] RN Upton and DR Mould. Basic Concepts in Population Modeling, Simulation, and Model-Based Drug Development: Part 3: Introduction to Pharmacodynamic Modeling Methods. *CPT: Pharmacometrics & Systems Pharmacology*, 3(1):88, January 2014.