ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

Symbolic Regression for Approximating Analytic Solutions to Differential Equations

Dionisel Y. Regalado

North western Mindanao State College of Science and Technology, Philippines

DOI: https://doi.org/10.51244/IJRSI.2025.120800044

Received: 24 July 2025; Accepted: 31 July 2025; Published: 02 September 2025

ABSTRACT

Approximate analytic expressions are obtained for initial value problems with purely numerical solutions. Symbolic regression was utilized to obtain such analytic expression. For functions that are Lipschitz continuous, results revealed that the maximum absolute error (sup-norm) is bounded.

Keywords: symbolic regression, initial value problem, finite difference method

INTRODUCTION

Many differential equations used in Engineering and the Sciences have no closed-form analytic solutions but are numerically solved (Hoffman et al., 2001). Numerical methods are techniques used to evaluate a solution at a point without necessarily knowing the exact analytic form of the solution to an ordinary differential equation (ODE). In most application, approximate numerical solutions may be sufficient but for purposes of analyzing the mathematical properties of the solution, it may be necessary to derive an analytic expression for it. For such situations, approximate symbolic expressions may serve as important bases for further mathematical analysis. Recent developments in genetic programming can be exploited to provide such approximate solutions through symbolic regression.

A first order differential equation is an initial value problem (IVP) if (Iserlas, 2008):

$$u'(t) = f\big(t,u(t)\big),\ u(t_0) = u_0 \qquad (1)$$

where $f : [t_0,\infty) x R^d \to R^d$ and $u_0 \in R^d$. For higher order systems, it is possible to analyze the system as a larger set of first order systems such as (1) by employing extra variables. Thus, without loss of generality, one may restrict attention to (1) since a higher order system can be converted to a larger system of first order systems. For instance,

$$u'' + u = 0 \Leftrightarrow u' = z \text{ and } z' = -u \qquad (2)$$

where z is an extra variable.

Existence of a unique solution to (1) is guaranteed by the Picard-Lindelöf Theorem which states that a unique solution exists provided f is Lipschitz continuous. We state Lipschitz continuity in the case of real-valued functions. A real-valued function $f : R \to R$ is Lipschitz continuous if there exists a positive real K such that for all real x_1 and x_2:

$$|f(x_1) - f(x_2)| \leq K|x_1 - x_2| \qquad (3)$$

In fact, let $g : R \to R$ be everywhere differentiable, then g is Lipschitz continuous if and only if g has a bounded first derivative (Grossman et al.,2007).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

A numerical method for solving (1) with boundary conditions is the finite difference method. Starting from (1), replace the derivative $u'(t)$ by the approximation (Strikwerda, 2004):

$$u'(t) \approx \frac{u(t+h)-u(t)}{h} \tag{4}$$

which gives:

$$u'(t) \approx u(t) + hf(t,u(t)) \tag{5}$$

To apply formula (5), choose a step-size h and construct the sequence t_0, $t_1 = t_0 + h$, $t_2 = t_0 + 2h$, …, $t_n = t$ and denote by: $u_n = u(t_n)$.

Expression (5) is then transformed into recursive relation:

$$u_{n+1} = u_n + hf(t_n, u_n) \tag{6}$$

By choosing h small enough, we obtain the ordered pairs $\{(t_0,u_0),(t_1,u_1),(t_2,u_2),...,(t_n,u_n)\}$. We now seek a function $u*(t)$ that passes through all of the points such that:

$$\varepsilon = \sup_{t_i} |u(t) - u*(t)| \tag{7}$$

is minimum.

Theoretical and Experimental Results

We start by defining an initial value problem.

Definition 1. Let $f: [t_0, \infty) x\ \mathbb{R} \to \mathbb{R}$ be a real – valued function. A first order equations is an initial value problem (IVP) if

$$U'(t) = f(t, u(t)), \quad u(0) = u_0 \in \mathbb{R}.$$

The function $u(t)$ is a solution to the IVP if it satisfies the differential equation. We seek necessary and sufficient conditions for the IVP to have a unique solution.

Definition 2. A real-valued function $f: \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous if there exists a constant $k > 0$ such that for all $x_1, x_2 \in \mathbb{R}$,

$$|f(x_1) - f(x_2)| \leq k|x_1 - x_2|.$$

Lemma 1. Let f be Lipschitz continuous, then f has a bounded first derivative.

Proof: From the definition of a derivative:

$$\frac{df}{dt} = \lim_{\Delta t \to 0} \frac{f(t+\Delta t) - f(t)}{\Delta t}.$$

By Lipschitz continuity of f, there exists as $k > 0$ for which

$$|f(t + \Delta t) - f(t)| \leq k|\Delta t|, \forall t$$

which yields

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

$$\frac{df}{dt} = \lim_{\Delta t \to 0} \frac{f(t+\Delta t) - f(t)}{\Delta t} \leq \lim_{\Delta t \to 0} \frac{k\Delta t}{\Delta t} = k.$$

It follows that

$$\frac{df}{dt} \leq k \qquad \Delta t \to 0. \qquad \blacksquare$$

The Picard – Lindelof Theorem guarantees a unique solution to the IVP.

Theorem 1. $(Picard - Lindelöf)$. Let

$$u'(t) = f\big(t, u(t)\big), \quad u(t_0) = u_0.$$

Let f be uniformly Lipschitz continuous in u and continuous in t. then for a given $\varepsilon > 0$, there exists a unique solution to the IVP on $[t_0 - \varepsilon, t_0 + \varepsilon]$.

Proof. Write the IVP as the integral equation:

$$u(t_0) = u_0 + \int_0^t f\big(s, u(s)\big)ds \quad \forall t \in [t_0, t_0 + \varepsilon]. \tag{8}$$

where ε is to be determined. Define:

$$T: C[t_0,\ t_0 + \varepsilon] \to C[t_0,\ t_0 + \varepsilon]$$

$$T\big(u(t)\big) = u_0 + \int_0^t f\big(s, u(s)\big)ds$$

Hence, (8) is a fixed point of T. we show that T satisfies a Lipschitz condition:

$$\begin{aligned} \|T(x)(t) - T(y)(t)\| &= \left\| \int_{t_0}^t f\big(s, x(s)\big)ds - \int_{t_0}^t f\big(s, y(s)\big)ds \right\| \\ &\leq \int_{t_0}^t \left\| f\big(s, x(s)\big) - f\big(s, y(s)\big) \right\| ds \\ &\leq \int_{t_0}^t M\|x(s) - y(s)\|ds \quad \text{where M is independent of u by Uniform continuity} \\ &\leq (t - t_0)M\|x - y\|_\infty \\ &\leq \varepsilon\, M\|x - y\|_\infty \end{aligned} \tag{9}$$

where the norm on $C[t_0,\ t_0 + \varepsilon]$ is:

$$\|u - y\|_\infty = \sup_t \{u(t) - y(t)\}.$$

Choose $\varepsilon < \frac{1}{M}$. It follows from this choice of ε and (9) that T is a contraction mapping and is Lipschitz continuous. By Banach's fixed point theorem, there exists a unique fixed point $u(t)$ for which:

$$T(u)(t) = u(t)$$

which solves the IVP. $\blacksquare$

Finite Difference Method.

Consider the IVP and suppose that we wish to find $u(t_0) = u_b$ at same later value t_b. If a closed form analytic solution can be obtained, then the problem is trivial. If, however, no such closed – form solution is available,

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

then we turn to numerical methods (Strikwerda, 2004). The simplest and most commonly-used method is the finite difference method. The method is based on the Taylor series expansion:

$$u(t+h) = u(t) + \frac{u'(t)h}{1!} + \frac{u''(t)h^2}{2!} + \cdots \quad (10)$$

By truncating all terms after the first derivative term, we have:

$$u(t+h) = u(t) + u'(t)h+) + O(h^2) \quad (11)$$

Definition 4. A function $f(t)$ is "big Oh" $g(t)$, written,

$$f(t) = O(g(t))$$

If there exists a constant $M > 0$ such that

$$f(t) = M|g(t)| \text{ as } t \to \infty$$

or:

$$\lim_{t\to\infty} sup_t \frac{f(t)}{g(t)} = M \quad \blacksquare$$

The error term in (11) is $O(h^2)$ which tends to zero as $h \to 0$. It follows that:

$$u(t+h) \backsimeq u(t) + u'(t)h \quad (12)$$

with an error proportional to h^2. Equation (12) can be rewritten as a recursive relation:

$$u(ih) = u\big((i-1)h\big) + u'\big((i-1)h\big)h \quad (13)$$

$$u_n = u_{n-1} + u'_{n-1}\, h$$

The closed interval $[t_0,\ t_b\ ]$ is divided into non-overlapping sub-intervals of length h, i.e. $|[t_i,\ t_{i-1}\ ]| = h$. The number of such sub-intervals is:

$$n = \frac{t_b - t_0}{h} \quad (14)$$

Convergence. For (13) to be useful, we need to show that the sequence $\{u_n\}_{n=0}^{\infty}$ converges.

Lemma 2. For $t \in [t_0,\ t_0 + \varepsilon]$, the sequence $\{u_n\}_{n=0}^{\infty}$ of (13) is a Cauchy sequence.

Proof.

Let $n > m, m, n \in Z^+$ and let $u_n(t) = u_{n-1}(t) + f(t, u_{n-1})h$ where h is the step size and the function f is Lipschitz continuous. Now,

$$\begin{aligned} \|u_n - u_m\| &= \|u_n - u_{n-1} + u_{n-1} + \cdots + u_{m+1} - u_m\| \\ &\le \|u_n - u_{n-1}\| + \|u_{n-1} - u_{n-2}\| + \cdots + \|u_{m+1} - u_m\| \\ &\le \|f(t, u_{n-1})h\| + \|f(t, u_{n-2})h\| + \cdots + \|f(t, u_m)h\| \\ &\le Mh(\|(u_{n-1})\| + \|(u_{n-2})\| + \cdots + \|(u_m)\|) \end{aligned} \quad (15)$$

where M is a Lipschitz constant. Let

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

$$L = \max_k\{\|u_{n-k}\|\} \quad (16)$$

Then:

$$\|u_n - u_m\| \leq MLn\ h.\ \text{As}\ \ n \to \infty, h = O(\tfrac{1}{n^2}) \to 0, \quad (17)$$

hence,

$$\|u_n - u_m\| \to 0 \quad \blacksquare \quad (18)$$

Theorem 2. The sequence $\{u_n\}_{n=0}^{\infty}$ converges.

Proof.

Since $\{u_n(t)\}_{n=0}^{\infty}$ is a Cauchy sequence of real numbers for each t, it follows that $\{u_n(t)\}$converges. The $Picard - Lindelöf$ theorem guarantees that $u_n(t) \to u(t)$ as $n \to \infty$ ■

Symbolic Regression

Symbolic regression is a type of regression analysis that does not specify the functional form of relationships between two variables. It utilizes a genetic algorithm to execute the analysis. This is included as an option in much statistical software. For this purpose, a one-month trial version of the software EUREQA was used to analyze the given data set (Regalado et al., 2019.

The data set analyzed consists of the ordered pairs $\{(t_i,u_i)\}_{i=0}^{n}$ generated by the recursive relation (13). As shown in the preceding sections, the computed values $u_i(t_i)$ can approximate the true solution $u\ (t_i)$ as closely as desired. These approximations provide a reliable input for symbolic regression, enabling the discovery of an analytic expression that best fits the numerical solution.

Let $t_i = ih, \quad i = 0,1,2, \dots, n$ and consider the pairs $\{(t_i, u_i)\}$. In traditional regression analysis, we assume a model of the form:

$$u(t) = g(t) + \varepsilon(t) \quad (19)$$

where $g(t)$ is a functional form that is completely specified except for the parameter values and $\varepsilon(t)$ are random errors with zero expectation and constant variance. For instance, g(t) may be a third degree polynomial:

$$u(t) = a + bt + ct^2 + dt^3 + \varepsilon(t) \quad (20)$$

where a, b, c and d are parameter to be estimated from the data.

In symbolic regression, the functional form of $g(t)$ is not specified but is assumed to be derived from a class of functions called building blocks. Let

$$\beta = \{g | g(t) \in C[t_0, t_0 + \varepsilon]\} \quad (21)$$

be the building blocks of function that are continuous. Symbolic regression, then, searches the space β for an optimal combination of building blocks that best fit the observations. Fitness is a user-defined quantity such as the mean – absolute error (MAE), the maximum error (ME) or the squared correlation goodness of fit (R^2).

The search process is implemented by applying the principles of genetic algorithm (GA). We consider the case when β has a finite number of building blocks:

$$\beta = \{g_1, g_2, \dots, g_m\}$$

Each $g_i(t)$ is assigned as fitness value e.g. maximum error, when fitted to the observations. Let

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

$$\Im = \{F(g_1), F(g_2), \dots, F(g_m)\} \quad (22)$$

be the set of fitness values for the building blocks. The building are then arranged from the fittest to the least – fit function. Let

$$\Im_{(I)} = \{F(g_{(1)}), F(g_{(2)}), \dots, F(g_{(m)})\}$$

where the subscripts denote ordering based on the fitness values i.e. $F(g_{(i)})$ is more fit than $F(g_{(i-1)})$.

A second generation of building blocks is obtained by combining the most fit building blocks and recomputing the fitness values of the resulting combinations of building blocks. The process continuous until a desired fitness value is obtained. Gelly et al. (2017) proved under sufficient conditions , the solution given by Genetic Programming converges when the number of examples goes to infinity , toward the actual function used to generate the examples. This property is known in Statistical Learning as **Universal Consistency**. The authors provided new results in the direction of bloat analysis: structural bloat and functional bloat. Structural bloat occurs when no optimal solution , that is, when no function exactly matches all possible examples, is approximated by the search space. In such cases, the authors demonstrated that optimal solutions of increasing accuracy will also exhibit increasing complexity.). On the other hand, **functional bloat** is defined as the bloat that takes place when programs length keeps on growing even though an optimal solution (of known complexity) does lie in the search space.

Illustration and Application

We provide a simple illustration of the proposed method for deriving an approximate analytic solution to a differential equation based on a numerically-derived ordered pairs. The illustration can be solved analytically. Thus, the solution is known which allows us to evaluate the proposed procedure more clearly.

$$The\ IVP: u'(t) = 5u + 2, \quad u(0) = 0$$

$$The\ analytic\ solution: u(t) = .4\ (\exp(5t) - 1)$$

We confine the search domain on the interval $t\ \varepsilon$ [0,1] and we seek the value u(1). We chose step sizes h = .1, .01, .001, .0001, .00001 and observed the maximum absolute error:

$$Sup = max_t \left| u(t) - u(t)_{pred} \right|$$

The finite difference equation gives

$$u_n = u_{n-1} + (5u_{n-1} + 2)h, \ u_0 = 0$$

Table 1 shows the relationship between step size and the maximum absolute error (sup).

Table 1: Relationship of step size and maximum absolute error

step size	sup
0.1	36.3
0.01	6.765
0.001	0.735
0.0001	0.0744
0.00001	0.0742

The estimated regression equation shows that as the step size increases, the maximum absolute error likewise increases.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

The regression equation is

$$sup\ 0.865 + 357\ step\ size$$

Predictor	Coef	SE Coef	T	P
Constant	0.8647	0.7886	1.10	0.353
step siz	356.63	17.55	20.33	0.000

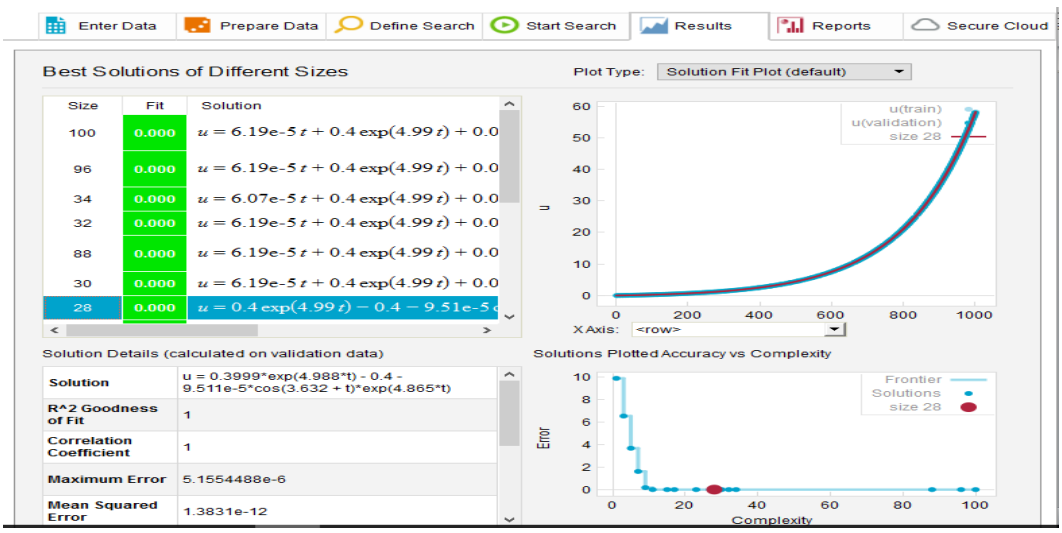
S = 1.533 R-Sq = 99.3% R-Sq(adj) = 99.0%

Using h = 0.001, we used symbolic regression using a trial version of EUREQA to obtain a closed form expression for the solution of the IVP. A portion of the data set is reproduced below:

Table 2: Portion of the Finite Difference Solution with h = .001

T	**u**
0	0
0.001	0.002
0.002	0.00401
0.003	0.00603
0.004	0.00806
0.005	0.010101
0.006	0.012151
0.007	0.014212
0.008	0.016283
0.009	0.018364
0.01	0.020456

The solution chosen has a maximum error of 0.00000515. The exact expression is shown on the screenshot below:



ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

Figure 1: Screenshot of the Final Solution

The final approximate solution to the IVP when rounded to the first decimal place is:

$$u(t) = .4(\exp(5t) - 1) - (9.5x\ 10^{-5})\cos(3.6 + t)\exp(5t)$$

The maximum difference between the approximate solution and the actual solution is:

$$\left|u(t) - \widehat{u(t)}\right| = .00155$$

Let

$\hat{u}(t_i)$ =solution obtained at t_i by the finite difference method

$\hat{\hat{u}}(t_i)$ = symbolic regression value at t_i using $\hat{u}(t_i)$ as inputs

$u(t_i)$ = actual value of the solution at t_i

Then

$$\sup_{0\le t_i\le t}\left\|u(t_i) - \hat{\hat{u}}(t_i)\right\| = \sup_{0\le t_i\le t}\left\|u(t_i) - \hat{u}(t_i) + \hat{u}(t_i) - \hat{\hat{u}}(t_i)\right\|$$

$$\le \|u(t_i) - \hat{u}(t_i)\|_\infty + \left\|\hat{u}(t_i) - \hat{\hat{u}}(t_i)\right\|_\infty$$

$$\le \varepsilon_1 + \varepsilon_2 = \varepsilon$$

As $h \to 0$, $\varepsilon_1 \to 0$ and if the solution $u(t)$ is in the search space $\beta, \varepsilon_2 \to 0$. Thus, $\left\|u(t_i) - \hat{\hat{u}}(t_i)\right\|_\infty \to 0$ as $n \to \infty$.

REFERENCES

1. Blickle T. and Thiele L. (1994).Genetic programming and redundancy. In J. Hopf, editor, Genetic Algorithms Workshop at KI-94, pages 33–38. Max-Planck-Institut f¨ur Informatik.
2. Gelly Sylvain, Teytaud Olivier, Bredeche Nicolas, Schoenauer Marc (2017) Symbolic regression, parsimony, and some theoretical considerations about GP search space **.**Equipe TAO - INRIA Futurs, LRI, Bat. 490, University Paris-Sud, 91405 Orsay Cedex. France
3. Grossmann, C, Hans-G. Roos; Martin Stynes (2007). Numerical Treatment of Partial Differential Equations. Springer Science & Business Media.
4. Hoffman JD; Frankel S (2001). Numerical methods for engineers and scientists. CRC Press, Boca Raton.
5. Iserlas, A. (2008). A first course in the numerical analysis of differential equations. Cambridge University Press.
6. Koza J. R.(1992). Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA, USA.
7. Regalado et al. (2019). Approximate Analytic Solution To The Lotka-Volterra Predator–Prey Differential Equations Model. Journal of Higher Education Research Disciplines 4 (1), 38-47.
8. Strikwerda, J/ (2004). Finite Difference Schemes and Partial Differential Equations (2nd ed.). SIAM.