Symbolic Regression Using α, β Operators and Estimation of Distribution Algorithms: Preliminary Results

Luis M. Torres-Treviño
Facultad de Ingeniería Mecánica y Eléctrica
Centro de Innovación e Investigación y Desarrollo en Ingeniería y Tecnología
Universidad Autonoma de Nuevo León
Avenida Universidad S/N
San Nicolás de los Garza, Nuevo León
luis.torrestv@uanl.edu.mx

ABSTRACT

Modeling processes is an important task in engineering; however, the generation of models using only experimental data is not a straightforward problem. Linear regression, neural networks, and other approaches have been used for this purpose; nevertheless, a mathematical description is desirable specially when an optimization is required. Symbolic regression has been used for generating equations considering only experimental data. In this paper, two new operators are proposed to represent a mathematical model of a process. These operators simplified the way for representing equations making possible its use as a symbolic regression. The correct model is generated selecting the appropriate operators and parameters using an evolutionary algorithm like the estimation of distribution algorithms. As a preliminary results, three cases are used to illustrated the performance of the proposed approach. The results indicates that the use of these α, β operators are a promising way to apply symbolic regression to model complex process.

Categories and Subject Descriptors

I.2.2 [**Artificial Intelligence**]: Automatic Programming-Program synthesismeasures, performance measures]

General Terms

Algorithms, Design, Experimentation

Keywords

Symbolic regression, α-β Operators, Estimation of distribution algorithms, Evonorm for Proceedings

1. INTRODUCTION

Symbolic regression has been used to represents models of complex processes where measured or experimental data is adjusted to a specific mathematical formulas. Usually analytic methods are used; however, the computer can be involved to automatizing this process using evolutionary computation. Two principal paradigms has been developed for symbolic regression: Genetic programming ([3, 4, 1]) and grammar evolution ([7, 10, 2])

In genetic programming, the mathematical formulas are presented by trees structures evaluated using Lisp programming language; however, the same representation is used considering other programming languages ([8]). Grammar evolution applies operators to a integer string to generate a program. The principal difference with genetic programming is that other programming language can be used not only a tree structure, so its not required a Lisp interpreter to evaluate the integer string, simplifying its implementation ([14]).

The idea of our proposal is to use multi-variable calculus ([11, 15, 9]) for optimization of industrial processes. The use of calculus explains why a mathematical model is preferred over other representations. Usually, a simple representation could be useful because the derivation of these equations could be easy. Given a process, a mathematical model is proposed; then, optimization approaches are applied to adjust some parameters of the proposed model and tested again measured data of the process. In some cases, a model is not provided so this one must be generated considering only experimental or measured data. A similar approach is proposed because a mathematical model is generated given a fixed structure (equation) formed by two proposed operators called α and β where the functions, variable interactions, and constants are changed by an evolutionary algorithm. The α and β operators make reference to a predefined mathematical functions. The α operator applies a mathematical function of one argument and the β operator requires two arguments and applies four basic mathematical operations of addition, substraction, multiplication and division. A modeling of a complex process can be made adjusting the parameters of these operators generating relative simple equations.

The methodology proposed, the α and β operators and the Evonorm algorithm are described in section II. In section III, a numeric experiments and results are presented considering measured data of three real process given in the literature. Duscussion, conclusion and future work is given in the last section.

GECCO'11, July 12–16, 2011, Dublin, Ireland.

Table 1: Parameters and mathematical functions of the α operator.

α operator	mathematical operation
1	$(k1x + k2)$
2	$(k1x + k2)^2$
3	$(k1x + k2)^3$
4	$(k1x + k2)^{-1}$
5	$(k1x + k2)^{-2}$
6	$(k1x + k2)^{-3}$
7	$(k1x + k2)^{1/2}$
8	$(k1x + k2)^{1/3}$
9	$exp(k1x + k2)$
10	$log(k1x + k2)$
11	$sin(k1x + k2)$
12	$cos(k1x + k2)$
13	$tan(k1x + k2)$

2. SYMBOLIC REGRESSION, α, β OPERATORS AND EVONORM

The modeling of processes usually requires a mathematical approach where the search of analogies and studies of similar processes are made. A second step imply the adjustment of parameters of a proposed mathematical model considering usually measured data. An experimental design is proposed in several cases when there are not available data; however, other problems are involved here because every experiment could be very expensive, so a low number of experiments is desirable specially when the process has several input variables and responses [6]. The α and β are defined below, then Evonorm algorithm is described and finally a methodology is proposed to use the α and β operators and Evonorm for symbolic regression considering only experimental data.

2.1 α and β operators

In this proposal, an equation is represented by the combination of two operators, the α operator and the β operator. An α operators is defines as a function that takes only one arguments and apply only one mathematical operation. Considering a review of several models of real process, 13 operations are defined (see Table 1). An α operator uses two parameters called $k1$ and $k2$ that are real numbers and an integer operator α that describe the mathematical operation. The operator α is defines as:

$$Opr_{\alpha}(x, k1, k2) = \alpha((k1 * x + k2)) \qquad (1)$$

where α={ X^1, X^2,X^3,X^{-1}, X^{-2}, X^{-3}, $X^{1/2}$, $X^{1/3}$, $exp(X)$, $log(X)$, $sin(X)$, $cos(X)$,$tan(X)$ }. Depending of the number selected, a specific operation is made, per example, if α=1 then the operation made is $(k1 * x + k2)$, if α= 13 then the operation made is $tan(k1 * x + k2)$. Every integer defines a mathematical operation described in Table 1.

A β operator is defined as a function that require two arguments and makes the four basic arithmetic operations β={+. −, ∗, / } so a β operator equal to 1 imply the plus operator or $\beta(a, b) = a + b$, and $\beta(a, b) = a/b$ if β=4. An hypothesis is presented: The use of α and β operators can be used to approximate non linear functions if a correct selection of its parameters is made. The problem is the correct selection of the operations and the parameters involved. The solution of this problem is made using an evolutionary algorithm, specifically an estimation of distribution algorithms. In this work, Evonorm is used to solve the problem of selection the suitable parameters (k's) and integers to define the α and β operations.

2.2 Evolutionary algorithm Evonorm

Evonorm is an easy way to implement an estimation of distribution algorithm [12, 13]. As a evolutionary algorithm selection of new individuals and the generation of a new population is used; however, the crossover and mutation mechanism is substituted by an estimation of parameters of a normal distribution function. The following steps are used in Evonorm:

1. Evaluation of a population P.
2. Deterministic selection of individuals from P to PS.
3. Generation of a new population using PS

A population P is a matrix of size I_p (total of individuals) and D_r (total of decision variables). A solution is a set of decision variables and this set is represented as a real vector. Every row of the population P represents a set of parameter of the solution. The selection mechanism is deterministic because the most fittest individuals are selected. Usually the number of selected individuals are lower than the number of the original population, usually a twenty or ten percent of the total population are selected. A random variable with normal distribution is estimated per decision variable, so a marginal distribution function is used. Two parameters are estimated, the mean and the standard deviation, that is determined using the values of the selected individuals. The population of selected individuals is a matrix Ps of size I_s (total of individuals selected) and D_r The equations (2, 3) are used to calculate the mean and standard deviation considering every vector of the population Ps.

$$\mu_{pr} = \sum_{k=1}^{I_s} (Ps_{pr,k})/I_s \qquad (2)$$

$$\sigma_{pr} = \sqrt{(} \sum_{k=1}^{I_s} (Ps_{pr,k} - \mu_{pr}))/I_s \qquad (3)$$

A new population is generated using the estimated normal random variables. This is a stocasting process;, however, an heuristic is used to maintain an equilibrium between exploration and exploitation, so new solutions can be found not necessary near of the mean calculated. The best solution found Ix at the moment is involved in the generation so in the 50% percent of the times the mean is used in the calculations and in the other 50% percent of the time the best solution found Ix is used as a mean as is shown in the following seudocode:

```
for k = 1 to I_p
 for pr = 1 to D_r
   if U() > 0.5
     P(k, pr) = N(μ_pr, σ_pr)
   otherwise
     P(k, pr) = N(Ix_pr, σ_pr)
```

end of condition
end of cycle pr
end of cycle k

The random variable $U()$ has a uniform distribution function, $N()$ is a random variable with a normal distribution function and D_r is the total of decision variables involved in the solution.

2.3 Methedology proposed

Considering a table of experimental data, it is possible to use the methodology proposed. This approach involves the following steps:

1. Selection of the number of the α and β operators.
2. Make a representation of the parameters k's, α and β operators for the evolutionary algorithm.
3. Execute a search algorithm to set the parameters considering several runs to minimize the difference between the real process and the model.
4. Decode the representation to generate the equations transforming the α and β operators to its corresponding mathematical operation.
5. Use the model for analysis, prediction, control or optimization.

The following cases illustrate the implementation of every step of the methodology proposed.

3. EXPERIMENTS AND RESULTS

An illustration of the proposed approach is made by symbolic regression of three processes. The experimental data was extracted from the book of Montgomery about regression [5]. The minimization of the mean square error is used as a fitness function for all the models proposed. A preprocessing is made in every experimental data to consider a normalization between 0 and 1. All the process uses the same algorithm Evonorm. This one uses 200 individuals, 50 ones are selected and the algorithm runs 200 generations.

The first process is a wind generator that relates the wind velocity and the DC generated. Table 2 illustrates the experimental data.

A configuration that uses three α operators per variable and two β operators is used. The representation used by the evolutionary algorithm requires six k parameters (real numers), three integers for the α operators and two integers for the β operators. Ten runs was made and the best solution was selected (Table 3). This solution has an mean square error of 0.1032508. The performance of the model is illustrated in figure 1.

The decoded representation generates equations (4-8). The $\alpha_1 = 3$ operator has the reference to the operation $(x * k11+, k21)^3$, the operation $\alpha_2 = 1$ is $(k11 * x + k21)$, $\alpha_3 = 5$ is $(x * k13+, k23)^{-2}$, and $\beta_1 = 1$ is the addition of the terms, and $\beta_2 = 3$ is the multiplication of the terms.

$$r1 = (x * k11 + k21)^3 \quad (4)$$

$$r2 = (x * k12 + k22) \quad (5)$$

$$r3 = (x * k13 + k23)^{-2} \quad (6)$$

$$r4 = r1 + r2 \quad (7)$$

$$y = r3 * r4 \quad (8)$$

Table 2: Experimental data of the first process ([5]).

Wind velocity (Mph)	DC generated
5.000	1.582
6.000	1.822
3.400	1.057
2.700	0.500
10.000	2.236
9.700	2.386
9.550	2.294
3.050	0.558
8.150	2.166
6.200	1.866
2.900	0.653
6.350	1.930
4.600	1.562
5.800	1.737
7.400	2.088
3.600	1.137
7.850	2.179
8.800	2.112
7.000	1.800
5.450	1.501
9.100	2.303
10.200	2.310
4.100	1.194
3.950	1.144
2.450	0.123

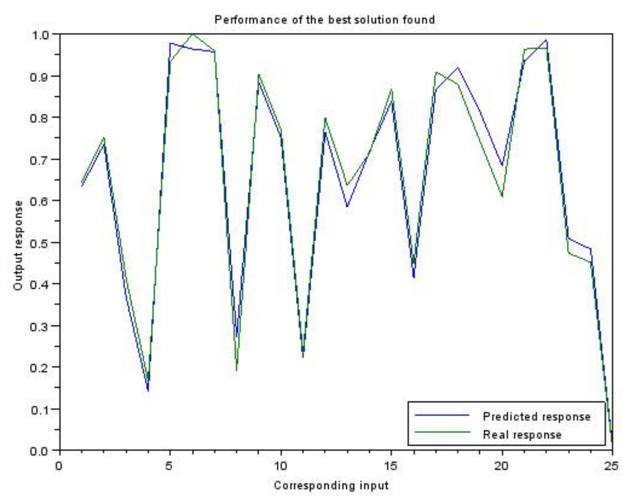

Figure 1: Performance of the model versus the performance of the real process.

Table 3: Results of the best solution found for the first process.

$\alpha - \beta$ operator	Value of the parameter
k11	0.7110103
k21	0.0047252
k12	0.9999239
k22	0.0072103
k13	0.6586252
k23	0.5209569
α_1	3
α_2	1
α_3	5
β_1	1
β_2	3

The equations can be resumes in the equation (9).

$$y = ((x*k11+k21)^3+(x*k12+k22))*(x*k13+k23)^{-2} \quad (9)$$

The second process is a little more complex. The value of 2-methoxyethanol to 1,2-dimethoxyethane (dimensionless) and the temperature (Celsius grades) are related with the cinematic velocity (m^2/s). Table 4 illustrates the information of the process.

The process involves two input variables and one output response, so a new model is required. The number of α and β operators are incremented considering four and three operators respectively (10-16).

$$r1 = \alpha_1(x_1 * k11 + k21) \quad (10)$$
$$r2 = \alpha_2(x_2 * k12 + k22) \quad (11)$$
$$r3 = \alpha_3(x_1 * k13 + k23) \quad (12)$$
$$r4 = \alpha_4(x_2 * k14 + k24) \quad (13)$$
$$y5 = \beta(r1, r2) \quad (14)$$
$$y6 = \beta(r3, r4) \quad (15)$$
$$y = \beta(r5, r6) \quad (16)$$

The characteristics of Evonorm are the same as the first process. Ten runs was made and the best solution was selected (Table 5). This solution has an mean square error of 0.0671133. The performance of the model is illustrated in Figure 2.

The decoded representation generates the equations (17-23).

$$r1 = (x_1 * k11 + k21)^3 \quad (17)$$
$$r2 = (x_2 * k12 + k22)^3 \quad (18)$$
$$r3 = (x_1 * k13 + k23) \quad (19)$$
$$r4 = cos(x_2 * k14 + k24) \quad (20)$$
$$r5 = r2/r1 \quad (21)$$
$$r6 = r3 + r4 \quad (22)$$
$$y = r6/r5 \quad (23)$$

The third process is a thermal solar energy system. It is desirable to model the heat flux (kWatts) considering five input variables, the sunstroke x_1 ($watts/m^2$), focus position in east direction x_2 (inches), focus position in south direction

Table 4: Experimental data of the second process ([5]).

x1	x2	y
0.92	-10	3.13
0.92	0	2.43
0.92	10	1.94
0.92	20	1.59
0.92	30	1.33
0.92	40	1.13
0.92	50	0.97
0.92	60	0.85
0.92	70	0.75
0.92	80	0.67
0.75	-10	2.27
0.75	0	1.82
0.75	10	1.49
0.75	20	1.25
0.75	30	1.06
0.75	40	0.92
0.75	50	0.8
0.75	60	0.71
0.75	70	0.63
0.75	80	0.57
0.57	-10	1.59
0.57	0	1.32
0.57	10	1.12
0.57	20	0.96
0.57	30	0.83
0.57	40	0.73
0.57	50	0.65
0.57	60	0.58
0.57	70	0.52
0.57	80	0.47
0.36	-10	1.16
0.36	0	0.99
0.36	10	0.86
0.36	20	0.75
0.36	30	0.67
0.36	40	0.59
0.36	50	0.53
0.36	60	0.48
0.36	70	0.44
0.36	80	0.4

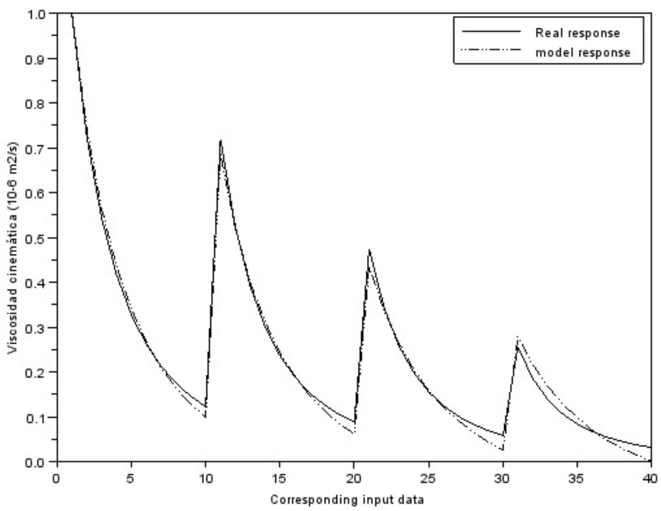

Figure 2: Performance of the model versus the performance of the real process.

Table 5: Results of the best solution found for the second process.

$\alpha - \beta$ operator	Value of the parameter
k11	0.3951954
k21	0.4894118
k12	0.6295830
k22	0.6217788
k13	0.9723338
k23	0.9600167
k14	0.0000054
k24	0.3142900
α_1	3
α_2	3
α_3	1
α_4	12
β_1	4
β_2	1
β_3	4

Table 6: Experimental data of the third process ([5]).

y	x1	x2	x3	x4	x5	
4540	2140	20640	30250	205	1732	99
4315	2016	20280	30010	195	1697	100
4095	1905	19860	29780	184	1662	97
3650	1675	18980	29330	164	1598	97
3200	1474	18100	28960	144	1541	97
4833	2239	20740	30083	216	1709	87
4617	2120	23305	29831	206	1669	87
4340	1990	19961	29604	196	1460	87
3820	1702	18916	29088	171	1572	85
3368	1487	18012	28675	149	1522	85
4445	2107	20520	30120	195	1740	101
4188	1973	20130	29920	190	1711	100
3981	1864	19780	29720	180	1682	100
3622	1674	19020	29370	161	1630	100
3125	1440	18030	28940	139	1572	101
4560	2165	20680	30160	208	1704	98
4340	2048	20340	29960	199	1679	96
4115	1916	19860	29710	187	1642	94
3630	1658	18950	29250	164	1576	94
3210	1489	18700	28890	145	1528	94
4330	2062	20500	30190	193	1748	101
4119	1929	20050	29960	183	1713	100
3891	1815	19680	29770	173	1684	100
3467	1595	18890	29360	153	1624	99
3045	1400	17870	28960	134	1569	100
4411	2047	20540	30160	193	1746	99
4203	1935	20160	29940	184	1714	99
3968	1807	19750	29760	173	1679	99
3531	1591	18890	29350	153	1621	99
3074	1388	17870	28910	133	1561	99
4350	2071	20460	30180	198	1729	102
4128	1944	20010	29940	186	1692	101
3940	1831	19640	29750	178	1667	101
3480	1612	18710	29360	156	1609	101
3064	1410	17780	28900	136	1552	101
4402	2066	20520	30170	197	1758	100
4180	1964	20150	29950	188	1729	99
3973	1835	19750	29740	178	1690	99
3530	1616	18850	29320	156	1616	99
3080	1407	17910	28910	137	1569	100

x_3 (inches), focus position in north direction x_4 (inches), and time (hours) of the day x_5.

The process involves five input variables and one output response. Only a basic configuration (one α operator per input variable) was used because the results indicates a good performance (24-32).

$$r1 = \alpha_1(x_1 * k11 + k21) \quad (24)$$
$$r2 = \alpha_2(x_2 * k12 + k22) \quad (25)$$
$$r3 = \alpha_3(x_3 * k13 + k23) \quad (26)$$
$$r4 = \alpha_4(x_4 * k14 + k24) \quad (27)$$
$$r5 = \alpha_5(x_5 * k15 + k25) \quad (28)$$
$$y1 = \beta_1(r1, r2) \quad (29)$$
$$y2 = \beta_2(r3, y1) \quad (30)$$
$$y3 = \beta_3(r4, y2) \quad (31)$$
$$y = \beta_4(r5, y3) \quad (32)$$

After ten runs of Evonorm algorithm, the best solution was selected and its results are shown in Table **??**. The performance of the model is illustrated in Figure 3.

The decoded representation generates the equations (33-41).

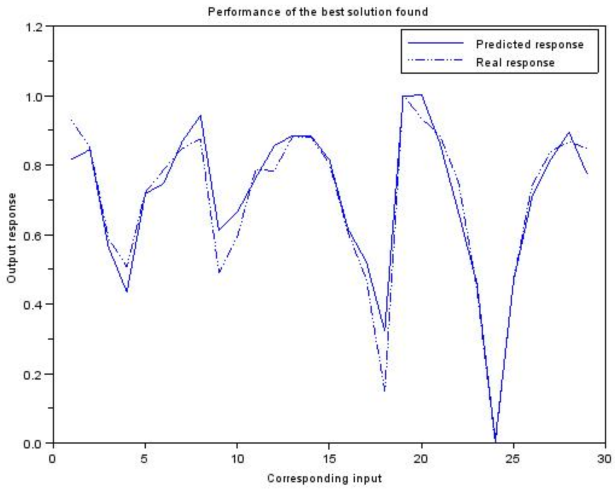

Figure 3: Performance of the best model found versus the real output response.

$$r1 = (x_1 * k11 + k21) \quad (33)$$
$$r2 = cos(x_2 * k12 + k22) \quad (34)$$
$$r3 = cos(x_3 * k13 + k23) \quad (35)$$
$$r4 = (x_4 * k14 + k24)^{-3} \quad (36)$$
$$r5 = (x_5 * k15 + k25) \quad (37)$$
$$y1 = r1/r2 \quad (38)$$
$$y2 = r3 * y1 \quad (39)$$
$$y3 = r6 - y2 \quad (40)$$
$$y = r5 * y3 \quad (41)$$

(**??**).

4. DISCUSSION, CONCLUSION AND FUTURE WORK

The easy implementation is the principal advantage of the proposed approach; however, several aspects about symbolic regression with these operators are open. Consider the following ten results to get a model of the first process (k constants were omitted for space, see Table 8):

Every result has a good performance, however, the mathematical operations are different in every model generated. Some solutions has complex mathematical functions, like the solution number 2, its requires two cosine functions. Other ones requires more simple functions like the solution number 10 because it requires only a root cubic operation and one multiplication. This variety of potential solutions open the opportunity to explore other models and consider other optimization criteria like multi - objective functions where a search of models with simple functions and good performance could be considered. Other way to generate a less complex equations is to establish other α operations or modify the set to exclude the cubic, root cubic and tan functions.

The proposed methodology implies the use of a predefined

Table 7: Results of the best solution found for the third process.

$\alpha - \beta$ operator	Value of the parameter
k11	0.9999917
k21	0.3199642
k12	0.9999974
k22	0.0270188
k13	0.4034232
k23	0.9474206
k14	0.2294144
k24	0.7506092
k15	0.1785192
k25	0.4411746
α_1	1
α_2	12
α_3	12
α_4	6
α_5	1
β_1	4
β_2	3
β_3	2
β_4	3
MSE	0.0035916

Table 8: Ten results for the first process.

Solution	α_1	α_2	α_3	β_1	β_2	MSE
1	14.	4.	1.	4.	1.	0.0100100
2	12.	12.	10.	3.	1.	0.0019435
3	8.	12.	7.	1.	3.	0.0014789
4	10.	3.	8.	4.	1.	0.0014698
5	8.	5.	7.	3.	3.	0.0018309
6	1.	12.	7.	1.	3.	0.0015609
7	11.	11.	1.	4.	4.	0.0019254
8	1.	12.	7.	1.	3.	0.0015742
9	1.	8.	1.	4.	4.	0.0015355
10	1.	1.	8.	3.	1.	0.0029486

configuration where α and β operators are included; however other configurations could be considered.

A demonstration about the universality approximation of the α-β operators must be considered because there are not a warranty that any configuration could be useful. The validation of every model is important for its usability, so a residual analysis must be considered; in resume, a regression suppositions must be considered in symbolic regression using α and β operators.

It is necessary to consider the performance of the proposed approach when more variables are involved. Usually three, four and more variables are involved so the number of the operators should be increased, however, this condition must be evaluated.

As a future work, the number of operations and the relationship with the number of variables will be evaluated to establish a requirement of an improvement. The assignation of the number of α and β operator must be automatized and be included in the representation to be used in the evolutionary algorithm. A theory must be established to demonstrate the hypothesis established above to consider the models of α and β operators as a universal approximator. Finally, the use of the equation for optimizing process will be considered as an application of this approach.

5. REFERENCES

[1] W. Banzhaf. *Genetic programming: First European workshop, EuroGP'98*. Springer, Paris, France, 1998.

[2] E. Hemberg, N. McPhee, M. O'Neill, and A. Brabazon. A. pre-, in- and postfix grammars for symbolic regression in grammatical evolution. In *IEEE Workshop and Summer School on Evolutionary Computing 18-22 IEEE Press*, 1998.

[3] J. R. Koza. *Genetic programming*. MIT Press, 1998.

[4] J. R. Koza, M. A. Keane, M. J. Streeter, W. Mydlowec, J. Yu, and G. Lanza. *Genetic programming IV: Routine Human-Competitive Machine Intelligence*. Kluwer Academic Publisher, 2003.

[5] D. C. Montgomery. *Introduction to Linear Regression Analysis*. John Wiley and Sons, 2007.

[6] D. C. Montgomery. *Design and analysis of experiments*. John Wiley and Sons, 2011.

[7] M. O'Neill and C.Ryan. Grammatical evolution: A steady state approach. In *Proceedings of the Second International Workshop on Frontiers in Evolutionary Algorithms.*, pages 419–423, 1998.

[8] E. V. Rick Riolo, Trent McConaghy. *Genetic Programming Theory and Practice VIII*. Springer, 2010.

[9] B. H. E. Ron Larson. *Multivariable Calculus*. Brooks/Cole, Cengage learning, 2009.

[10] C. Ryan, J. Collins, and M. O'Neill. Grammatical evolution: Evolving programs for an arbitrary language. In *First European Workshop on Genetic Programming. Lecture Notes in Computer Science 1391.*, 1998.

[11] J. Stewart. *Multivariable calculus; Concepts and contexts*. Thomson Brooks/Cole, 2005.

[12] L. Torres-T. Evonorm, a new evolutionary algorithm to continuous optimization. In *Workshop on Optimization by Building and Using Probabilistic Models (OBUPM-2006) Genetic and Evolutionary Computation Conference (GECCO-2006) CD Proceeding Tutorials and Workshops*, July 2006.

[13] L. Torres-T. Evonorm: Easy and effective implementation of estimation of distribution algorithms. *Journal of Research in Computing Science*, 23:75–83, November 2006.

[14] I. Zelinka, S. Celikovsky, and H. Richter. *Evolutionary Algorithms and Chaotic Systems*. Springer, 2010.

[15] D. G. Zill and W. S. Wright. *Multivariable Calculus*. Jones and Bartlett Publishers, 2011.