The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Racing Control Variable Genetic Programming for Symbolic Regression

Nan Jiang, Yexiang Xue (jiang631, yexiang@purdue.edu) Purdue University

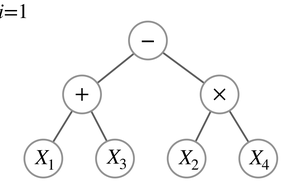
Symbolic Regression for Scientific Discovery

Given a dataset $D = [(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)]$, where $y_i \in \mathbb{R}$ and $\mathbf{x}_i = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^n$.
Find a closed-form equation ϕ that best fits the dataset D.

$$\phi^* \leftarrow \arg\min_{\phi \in \Pi} \frac{1}{N} \sum_{i=1}^{N} Loss(\phi(\mathbf{x}_i), y_i),$$

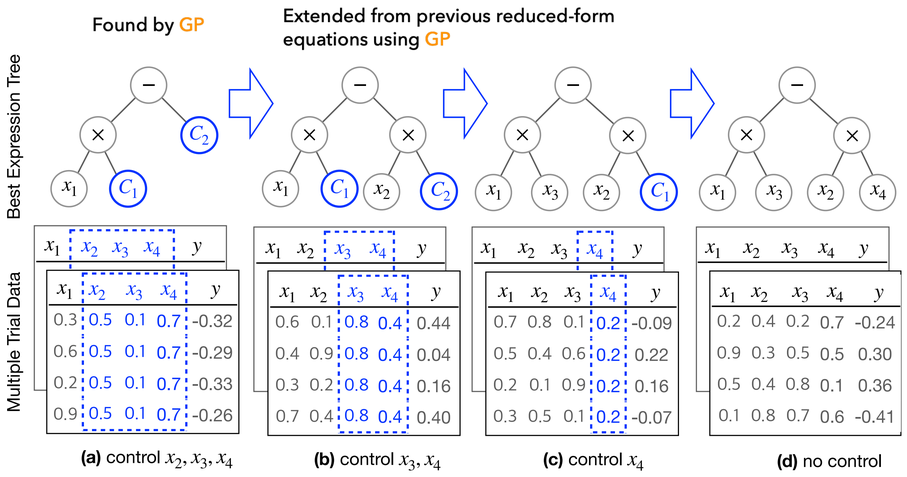
X_1	X_2	X_3	X_4	Y
0.3	0.5	0.1	0.7	-0.32
0.6	0.5	0.1	0.7	-0.29
0.2	0.5	0.1	0.7	-0.33
0.9	0.5	0.1	0.7	-0.26

(a) the dataset D



(b) The ground-truth expression $\phi = x_1 x_3 - x_2 x_4$

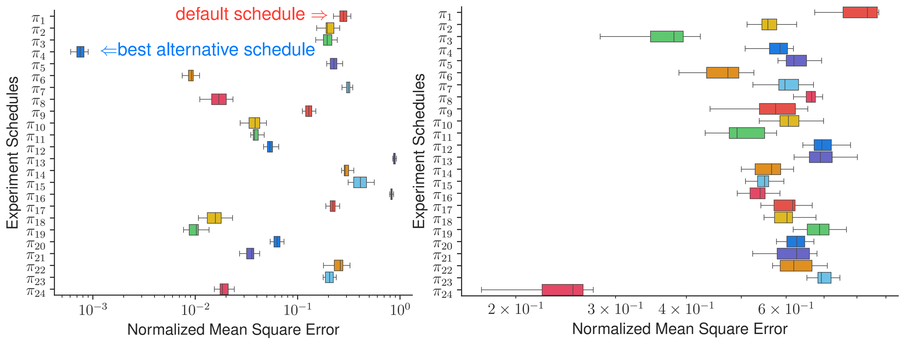
Background: Control Variable Genetic Programming (CVGP)



The data are generated from a data Oracle. The constants are identified using BFGS optimizer on batches of data

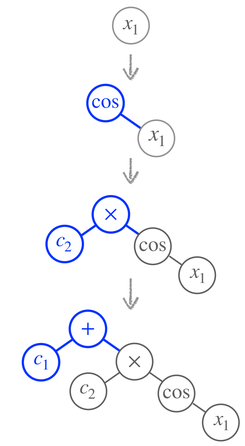
GAP: CVGP is sensitive to experiment schedule

There exists a better experiment schedule among all schedules than the default one (i.e., π_1), in terms of NMSE metric.

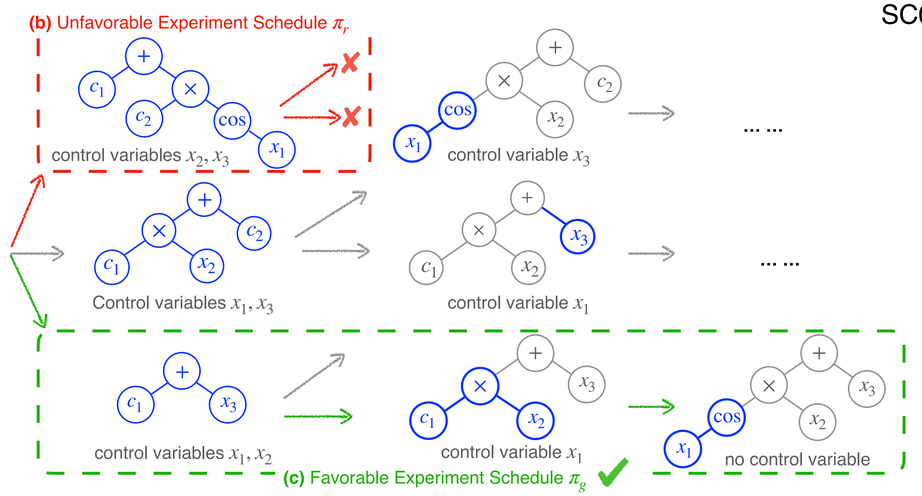


Our Method: Racing Control Variable Genetic Programming

(1) maintaining ***multiple*** experiment schedules rather than one.
(2) allowing **promising** experiment schedules to **survive** while letting **unfavorable** schedules **early stop**.



(a) Multi-steps of edits are needed to obtain the tree in (b).



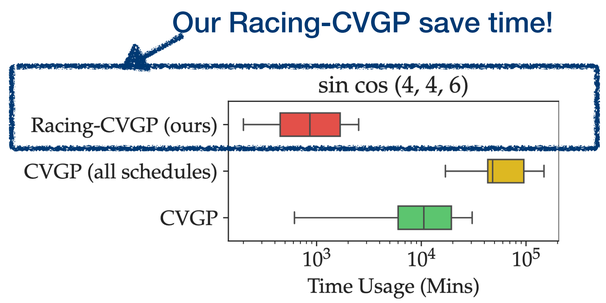
Running Time Analysis

The major hyper-parameters:

- the number of genetic operations per round, M;
- total rounds, n;
- the maximum size of population pool, N_p.

Our Racing-CVGP needs $\mathcal{O}(nMN_p)$, which is **roughly the same** as CVGP.

Experiment: Running-time Comparison



Experiment: Goodness-of-fit Benchmark

Our Racing-CVGP attains the smallest NMSE values.

Table 1: On Trigonometric datasets, median (50%) and 75%-quantile NMSE values of the expressions found by all the algorithms. Our Racing-CVGP finds symbolic expressions with the smallest NMSEs. "$T.O.$" implies the algorithm is timed out for 48 hours. The 3-tuples at the top $(\cdot, \cdot, \cdot)$ indicate the number of input variables, singular terms, and cross terms in the expression.

	(3, 2, 2)		(4, 4, 6)		(5, 5, 5)		(6, 6, 10)		(8, 8, 12)	
	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%
Racing-CVGP (ours)	**< 1E-6**	**< 1E-6**	**0.016**	**0.021**	**0.043**	**0.098**	**0.069**	**0.104**	**0.095**	**0.286**
CVGP	0.039	0.083	0.028	0.132	0.086	0.402	0.104	0.177	$T.O.$	$T.O.$
GP	0.043	0.551	0.044	0.106	0.063	0.232	0.159	0.230	$T.O.$	$T.O.$
Eureqa	**< 1E-6**	**< 1E-6**	0.024	0.122	0.158	0.377	0.910	1.927	0.162	2.223
DSR	0.227	7.856	2.815	9.958	2.558	3.313	6.121	16.32	0.335	0.410
PQT	0.855	2.885	2.381	13.84	2.168	2.679	5.750	16.29	0.232	0.313
VPG	0.233	0.400	2.990	11.32	1.903	2.780	3.857	19.82	0.451	0.529
GPMeld	0.944	1.263	1.670	2.697	1.501	2.295	7.393	21.71	$T.O.$	$T.O.$
SPL	0.010	0.011	0.144	0.231	0.147	0.280	0.472	0.627	0.599	0.746

Code & Acknowledgement

This research was supported by NSF grant CCF-1918327 and DOE – Fusion Energy Science grant: DE-SC0024583.

Our code implementation: