Control Variable Experiment for Symbolic Regression

Nan Jiang (jiang631@purdue.edu)
Department of Computer Science
Purdue University, USA

In collaboration with Yexiang Xue.

PURDUE
UNIVERSITY

Motivation: The history for the discovery of Ideal Gas Law

The task is to predict the governing equation among all the possible equations with the given dataset shown below, which is known as symbolic regression.

(a) The symbolic expression of ideal gas law.

Physical formula		Symbolic Expression	
$P = \frac{RnT}{V}$		$y = \frac{c_1 x_1 x_2}{x_3}$	
Symbols	**Physical Meaning**	**Variables**	**Variable Domains**
R	Ideal gas constant	Constant c_1	8.31446
n	Number of moles	Input variable x_1	$(0.01, 100)$
T	Absolute Temperature	Input variable x_2	$(0, 1000)$
V	Volume	Input variable x_3	$(0.001, 10)$
P	Absolute Pressure	Output variable y	

Current algorithms directly search for the optimal expression involving all three variables, which scales poorly to multiple-variable expressions.

(c) Dataset from the governing expression $y = c_1 x_1 x_2 / x_3$.

#Moles n(#mol)	Temperature T (Kelvin)	Volume V (m^3)	Pressure P (Pa)
Input variable x_1	Input variable x_2	Input variable x_3	Output y
0.58	291	0.002	6.90×10^5
44.50	273	1.00	1.01×10^5
10.00	273	1.00	2.27×10^4
…	…	…	…

Can we expediate the discovery process?

Motivation: The history for the discovery of Ideal Gas Law

- In 1663, Robert Boyle found

$$PV=\text{constant}$$

where the number of moles (n) and temperature (T) are fixed.

- In 1787 and again in 1802, Jacques Charles and Joseph Louis Gay-Lussac demonstrated

$$V/T = \text{constant}$$

where the number of moles (n) and pressure (P) are fixed.

- In 1811, Amedeo Avagadro demonstrated

$$V/n=\text{constant}$$

where the pressure (P) and temperature (T) are fixed.

- Finally, we arrived at the ideal gas law,

$$PV=nRT$$

The scientists use control variable experiment to solve a much simpler task.

Can we introduce this idea into symbolic regression, so that the algorithm mimic human scientist?

(a) The symbolic expression of ideal gas law.

Physical formula	**Symbolic Expression**
$P = \frac{RnT}{V}$	$y = \frac{c_1 x_1 x_2}{x_3}$

Symbols	**Physical Meaning**	**Variables**	**Variable**
R	Ideal gas constant	Constant c_1	8.31446
n	Number of moles	Input variable x_1	$(0.01, 100)$
T	Absolute Temperature	Input variable x_2	$(0, 1000)$
V	Volume	Input variable x_3	$(0.001, 10)$
P	Absolute Pressure	Output variable y	

Symbolic regression

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
3.0	-1.0	4.0	1
1.6	3.5	5.2	10.8
1.8	1.0	3.2	5
7.1	8.6	3.8	64.9
1.7	1.0	2.3	4
2.5	2.6	3.1	9.6
8.9	1.1	2.0	11.8
4.2	-1.0	2.2	-2
5.8	1.0	7.2	13
1.6	5.7	1.2	10.3
9.7	-1.0	1.7	-8

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?

Symbolic regression

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
1.8	1.0	3.2	5
1.7	1.0	2.3	4
5.8	1.0	7.2	13

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?
- How about if I only ask you to look into these rows?

$$y = x_1 + x_3?$$

Symbolic regression

X_1	X_2	X_3	Y
3.0	-1.0	4.0	1
4.2	-1.0	2.2	-2
9.7	-1.0	1.7	-8

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?
- How about if I only ask you to look into these rows?

$$y = x_1 + x_3?$$

- How about these rows?

$$y = -x_1 + x_3?$$

Symbolic regression

Red and blue data are two control variable experiment trials (X_2 controlled)!
Control variable experiments ***simplify*** symbolic regression!

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
3.0	-1.0	4.0	1
1.8	1.0	3.2	5
1.7	1.0	2.3	4
4.2	-1.0	2.2	-2
5.8	1.0	7.2	13
9.7	-1.0	1.7	-8

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?
- How about if I only ask you to look into these rows?

$$y = x_1 + x_3?$$

- How about these rows?

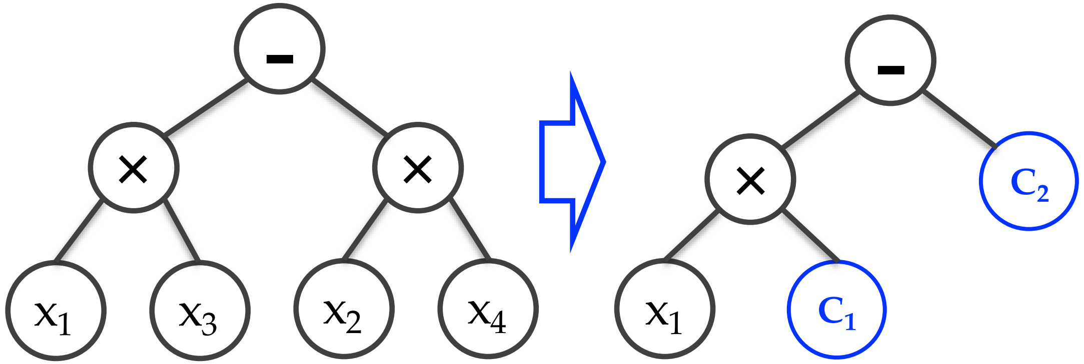
$$y = -x_1 + x_3?$$

- Maybe the equation is:

$$y = x_2 x_1 + x_3?$$

INDEED!

Control Variable Experiments



(a) Ground-truth expression

(b) Reduced form after controlling x_2, x_3, x_4

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

controlled (x_2, x_3, x_4)

(c) Trial T_1

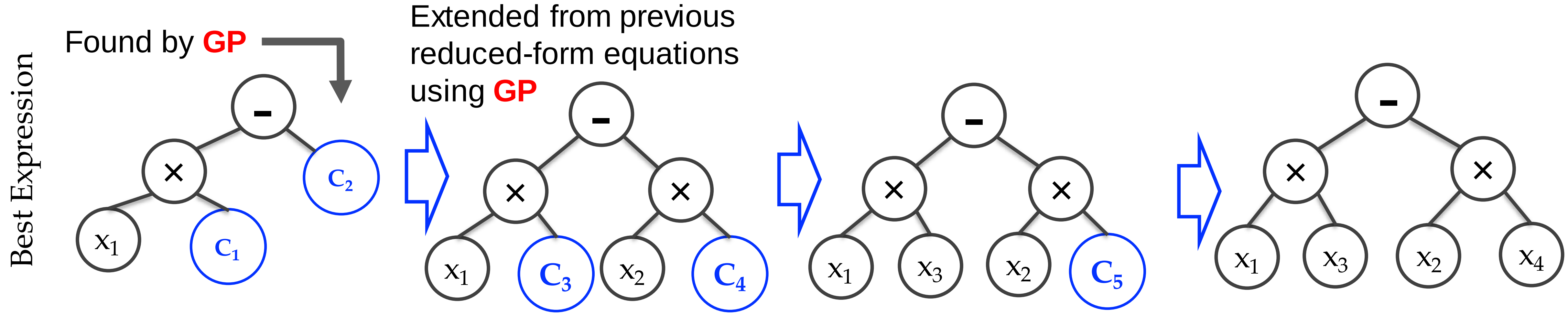
x_1	x_2	x_3	x_4	y
0.6	0.3	0.8	0.2	0.42
0.1	0.3	0.8	0.2	0.02
0.2	0.3	0.8	0.2	0.10
0.9	0.3	0.8	0.2	0.66

controlled (x_2, x_3, x_4)

(d) Trial T_2

- **Control variable experimentation** – a classic procedure widely implemented and proven useful in science.
- **Controlled variables**: take the same value in a trial, but vary in values across trials
- **Free variables**: values change within a trial
- **Ground-truth equation**: the hidden equation that generates the data
- **Reduced form equation**: Under a controlled experiment, the data looks “as if” generated by the reduced equation, in which controlled variables are replaced with constants.

Control Variable Experiment with Genetic Programming (CVGP)



Multiple Trials

x_1	x_2	x_3	x_4	y
0.3	0.5	0.1	0.7	-.32
0.6	0.5	0.1	0.7	-.29
0.2	0.5	0.1	0.7	-.33
0.9	0.5	0.1	0.7	-.26

(a) Control x_2,x_3,x_4

x_1	x_2	x_3	x_4	y
0.6	0.1	0.8	0.4	0.44
0.4	0.9	0.8	0.4	0.04
0.3	0.2	0.8	0.4	0.16
0.7	0.4	0.8	0.4	0.40

(b) Control x_3,x_4

x_1	x_2	x_3	x_4	y
0.7	0.8	0.1	0.2	-.09
0.5	0.4	0.6	0.2	0.22
0.2	0.1	0.9	0.2	0.16
0.3	0.5	0.1	0.2	-.07

(c) Control x_4

x_1	x_2	x_3	x_4	y
0.2	0.4	0.2	0.7	-.24
0.9	0.3	0.5	0.5	0.30
0.5	0.4	0.8	0.1	0.36
0.1	0.8	0.7	0.6	-.41

(d) No control

Experiment Results

Ops	Dataset configs	CVGP (ours) 50%	CVGP (ours) 75%	GP 50%	GP 75%	DSR 50%	DSR 75%	PQT 50%	PQT 75%	VPG 50%	VPG 75%	GPMeld 50%	GPMeld 75%
inv	(2,1,1)	0.198	0.490	**0.024**	**0.053**	0.032	3.048	0.029	0.953	0.041	0.678	0.387	22.806
	(4,4,6)	**0.036**	**0.088**	0.038	0.108	1.163	3.714	1.016	1.122	1.087	1.275	1.058	1.374
	(5,5,5)	0.076	0.126	**0.075**	**0.102**	1.028	2.270	1.983	4.637	1.075	2.811	1.479	2.855
	(5,5,8)	**0.061**	**0.118**	0.121	0.186	1.004	1.013	1.005	1.006	1.002	1.009	1.108	2.399
	(6,6,8)	**0.098**	**0.144**	0.104	0.167	1.006	1.027	1.006	1.020	1.009	1.066	1.035	2.671
	(6,6,10)	**0.055**	**0.097**	0.074	0.132	1.003	1.009	1.005	1.008	1.004	1.015	1.021	1.126
sin, cos	(3,2,2)	**0.098**	**0.165**	0.108	0.425	0.350	0.713	0.351	1.831	0.439	0.581	0.102	0.597
	(4,4,6)	**0.078**	**0.121**	0.120	0.305	7.056	16.321	5.093	19.429	2.458	13.762	2.225	3.754
	(5,5,5)	**0.067**	**0.230**	0.091	0.313	32.45	234.31	36.797	229.529	14.435	46.191	28.440	421.63
	(5,5,8)	**0.113**	**0.207**	0.119	0.388	195.22	573.33	449.83	565.69	206.06	629.41	363.79	666.57
	(6,6,8)	**0.170**	**0.481**	0.186	0.727	1.752	3.824	4.887	15.248	2.396	7.051	1.478	6.271
	(6,6,10)	**0.161**	**0.251**	0.312	0.342	11.678	26.941	5.667	24.042	7.398	25.156	11.513	28.439
sin, cos, inv	(3,2,2)	0.049	**0.113**	**0.023**	0.166	0.663	2.773	1.002	1.992	0.969	1.310	0.413	2.510
	(4,4,6)	**0.141**	**0.220**	0.238	0.662	1.031	1.051	1.297	1.463	1.051	1.774	1.093	1.769
	(5,5,5)	**0.157**	0.438	0.195	**0.337**	1.098	3.617	1.018	5.296	1.012	1.27	1.036	3.617
	(5,5,8)	**0.122**	**0.153**	[illegible]	0.186	1.009	1.103	1.017	1.429	1.007	1.132	1.07	2.904
	(6,6,8)	**0.209**	**0.590**	[illegible]	0.646	1.003	1.153	1.047	1.134	1.059	1.302	1.029	3.365
	(6,6,10)	0.139	0.232	[illegible]	**0.159**	1.654	3.408	1.027	1.069	1.009	1.654	1.445	2.106

Median (50%) and 75%-quantile NMSE values of the symbolic expressions found by all the algorithms on several noisy benchmark datasets. Our CVGP finds symbolic expressions with the smallest NMSEs.

Conclusions

- Control Variable Genetic Programming (CVGP) for symbolic regression
  - Learning from control variable experiments
  - Incrementally build complex equations from simple ones using genetic programming
- Look into future: passive learning vs. active probing
  - Science progress resulted from insightful experiment design, courageous hypothesis forming (reasoning) + high-capacity modeling (learning)