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

Racing Control Variable Genetic Programming for Symbolic Regression

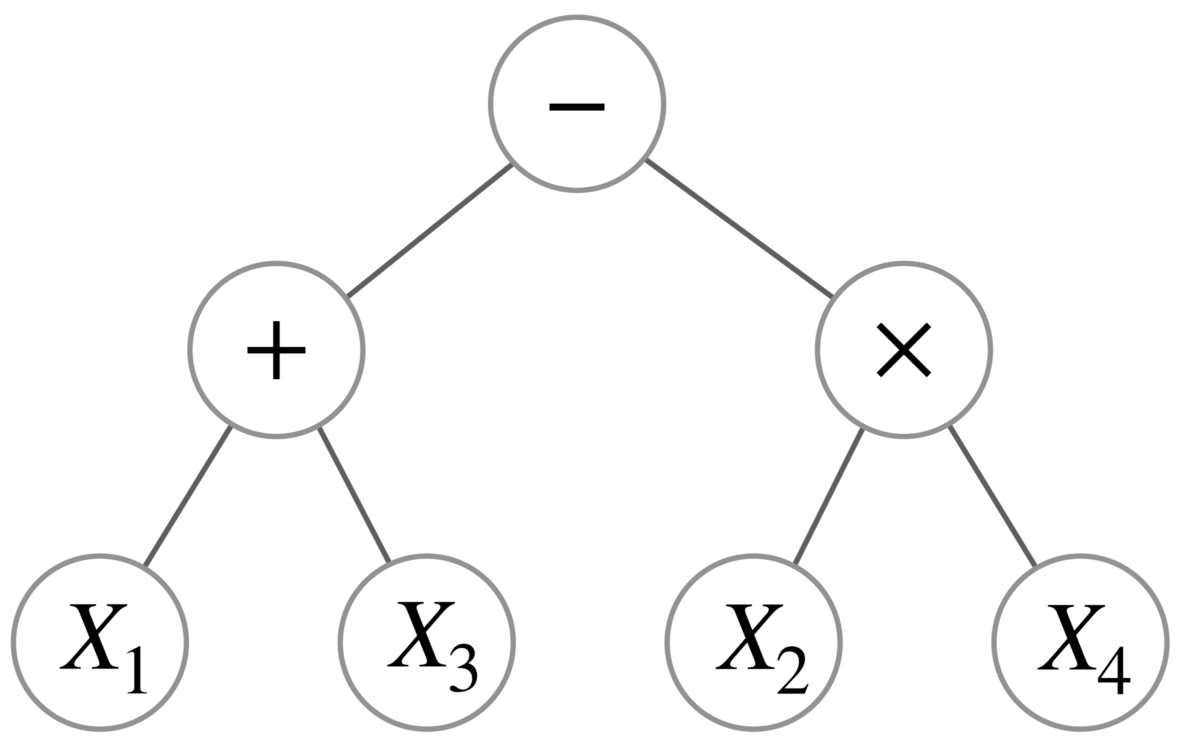
Nan Jiang and Yexiang Xue
Purdue University
jiang631@purdue.edu, yexiang@purdue.edu
Jan 2024

What is Symbolic Regression?

- 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.
- $$\phi^* \leftarrow \arg\min_{\phi \in \Pi} \frac{1}{N} \sum_{i=1}^{N} Loss(\phi(\mathbf{x}_i), y_i),$$
- Here is an example:

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) Ground-truth expression $\phi = x_1 x_3 - x_2 x_4$

Current Approaches and Challenges

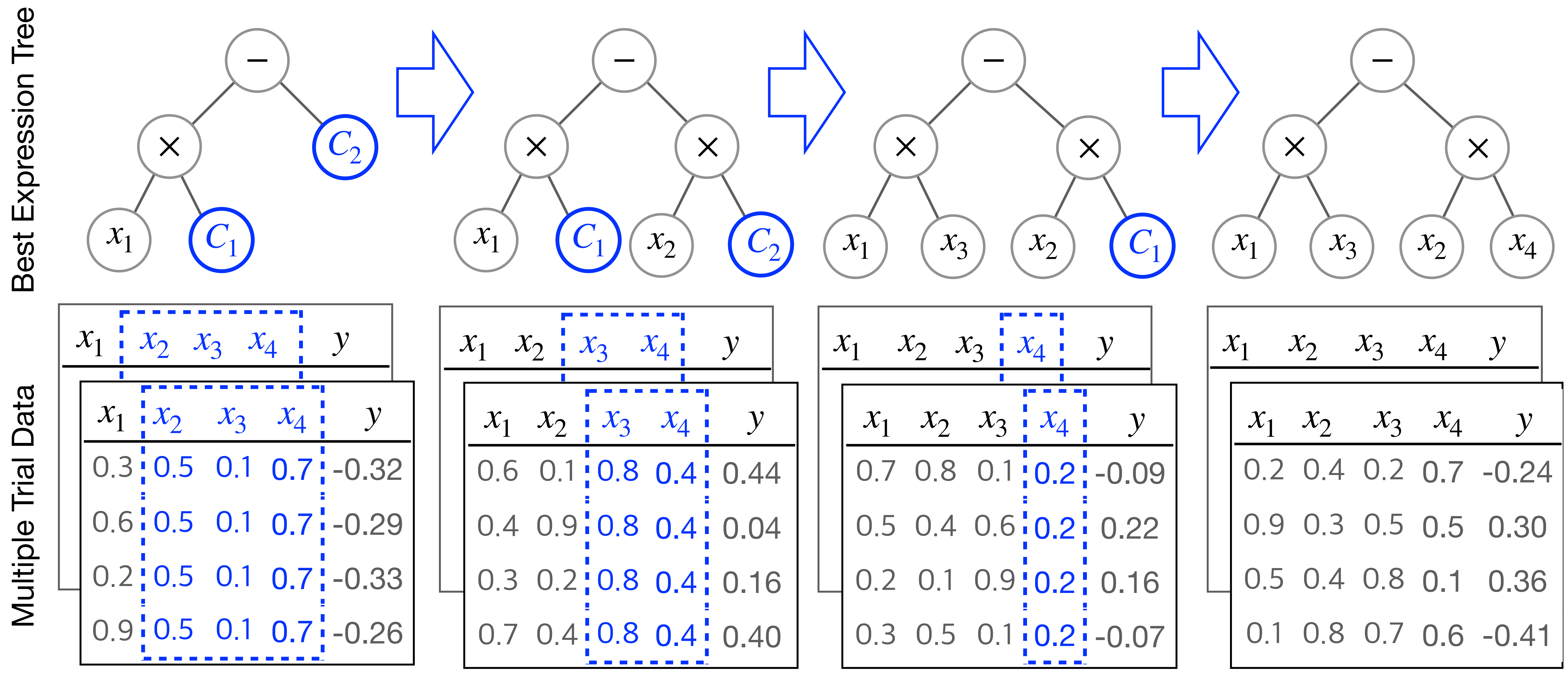
- Generic programming (GP): learn to find a equation with mutations and matings.
- Reinforcement learning (RL), Monte Carlo Tree Search (MCTS): learns to search a equation.
- Deep neural models: learn representation of all similar equations.

Current Challenges:

- The hypothesis space of candidate equations is exponential to the number of input variables.
- Current methods is only applicable to problems with a few variables.

Control Variable Genetic Programming (CVGP)
(ECML 2023)

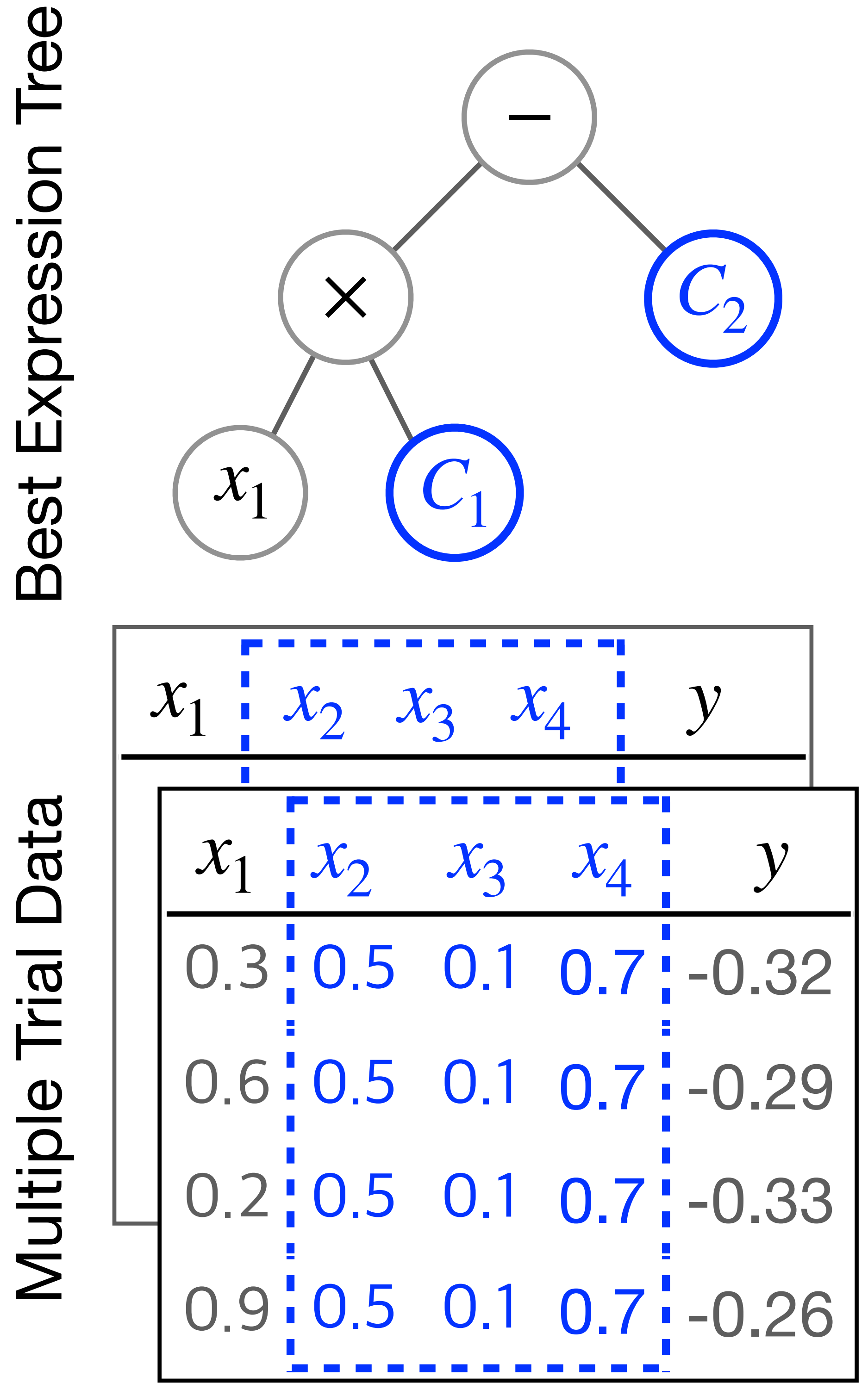
- Build the expression from simple to complex.



(a) control x_2, x_3, x_4 **(b)** control x_3, x_4 **(c)** control x_4 **(d)** no control

Control Variable Genetic Programming

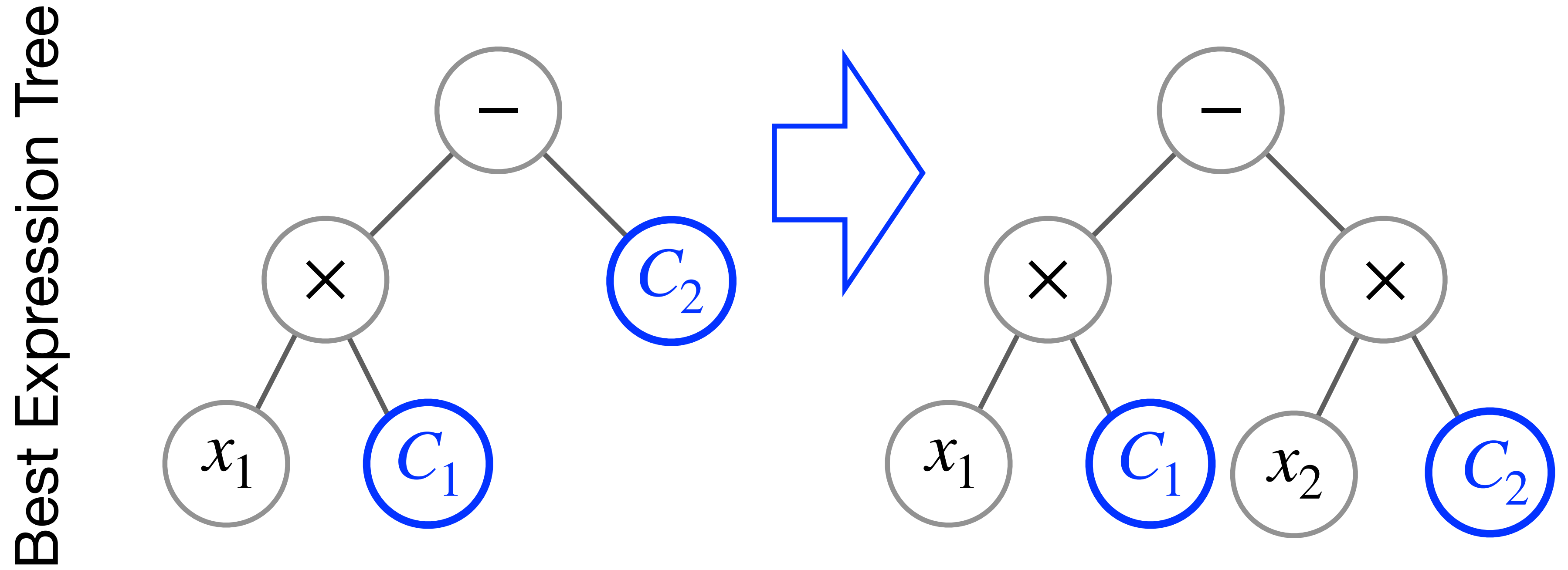
- Assumption: need a data oracle that can return the controlled variables dataset
- We can iteratively reduce the number of controlled variables.



(a) control x_2, x_3, x_4

Control Variable Genetic Programming

- Assumption: need a data oracle that can return the controlled variables dataset
- We can iteratively reduce the number of controlled variables.



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

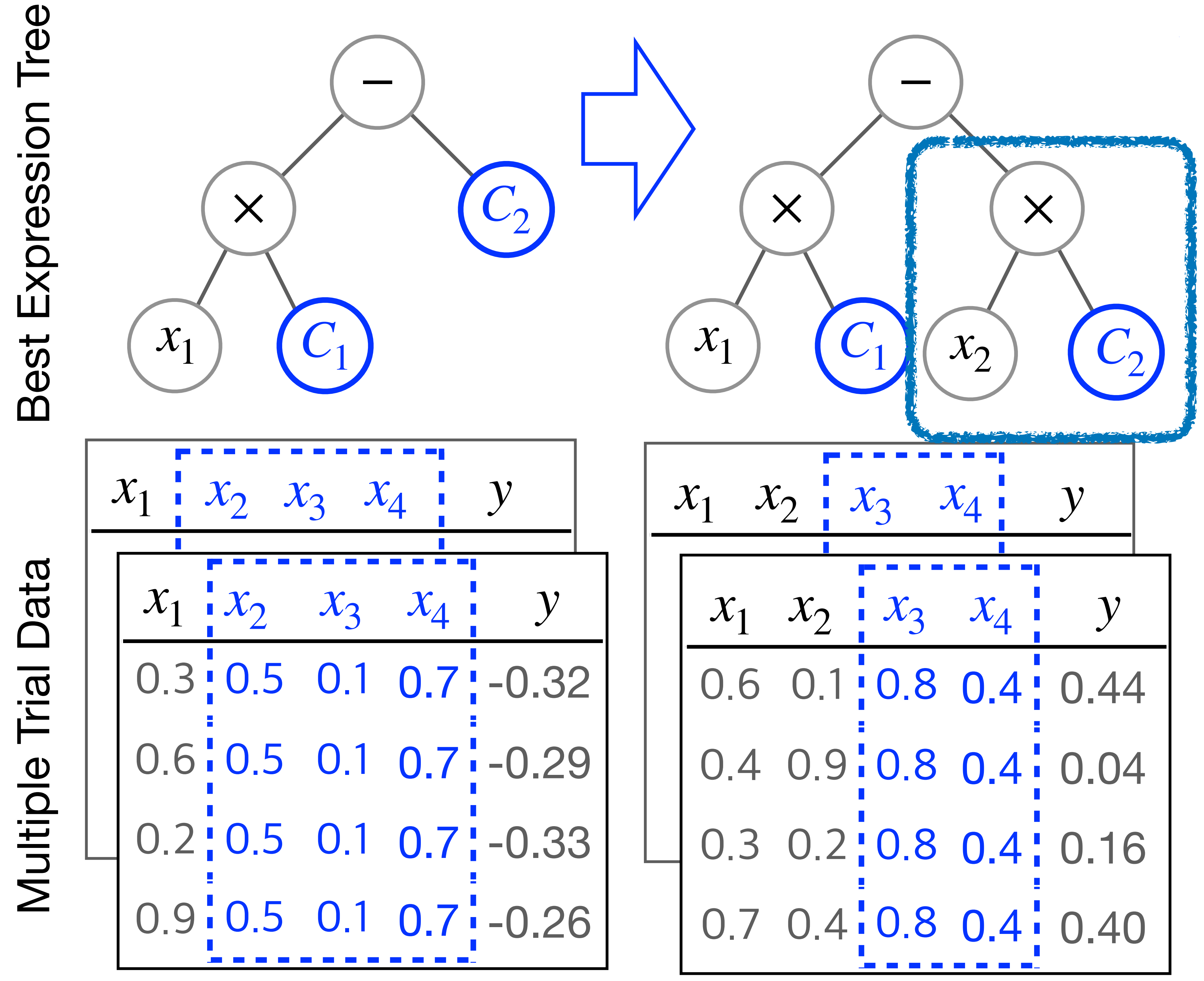
(a) control x_2, x_3, x_4

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

Control Variable Genetic Programming

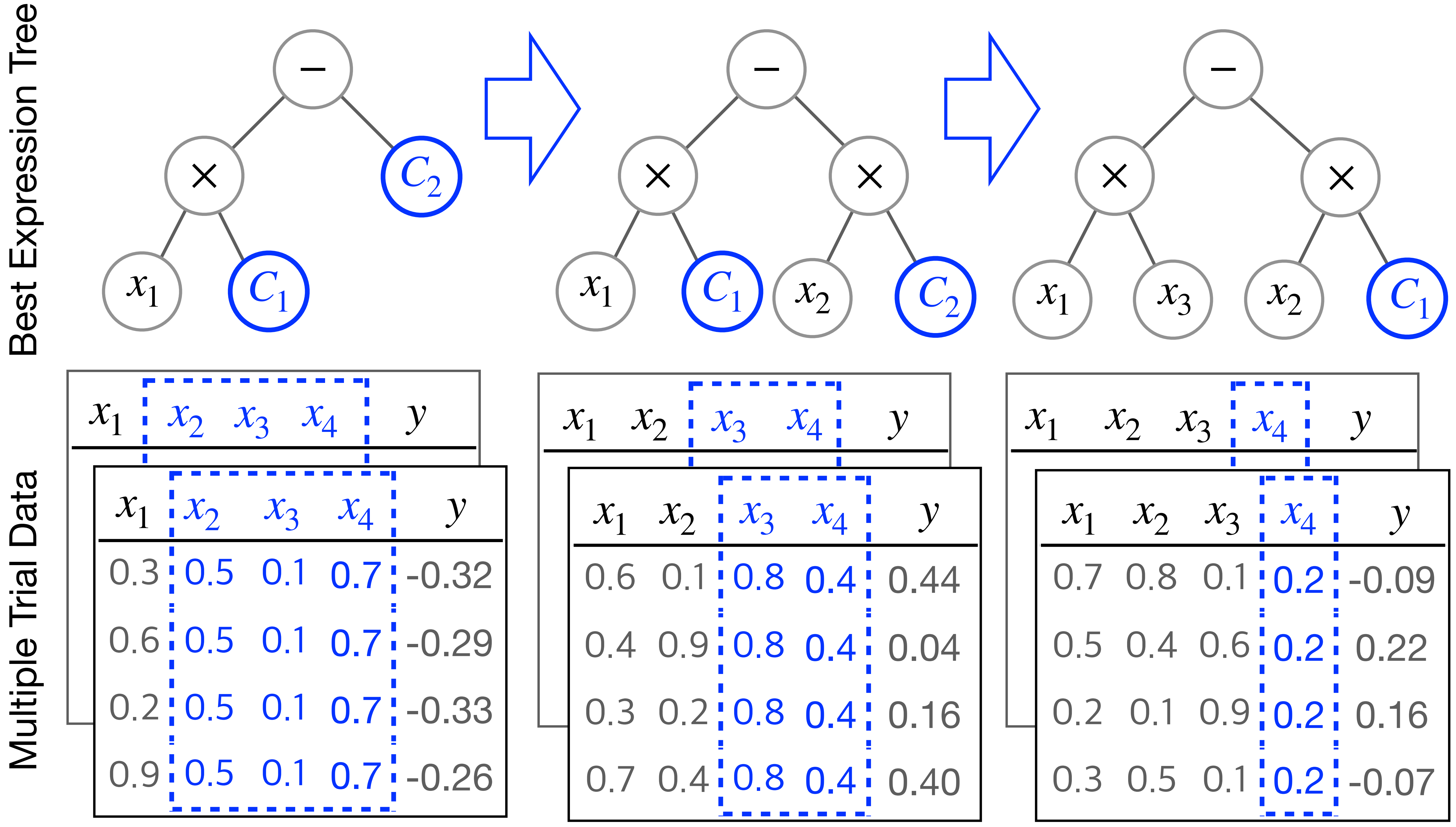
- Assumption: need a data oracle that can return the controlled variables dataset
- We can iteratively reduce the number of controlled variables.



(a) control x_2, x_3, x_4 **(b)** control x_3, x_4

Control Variable Genetic Programming

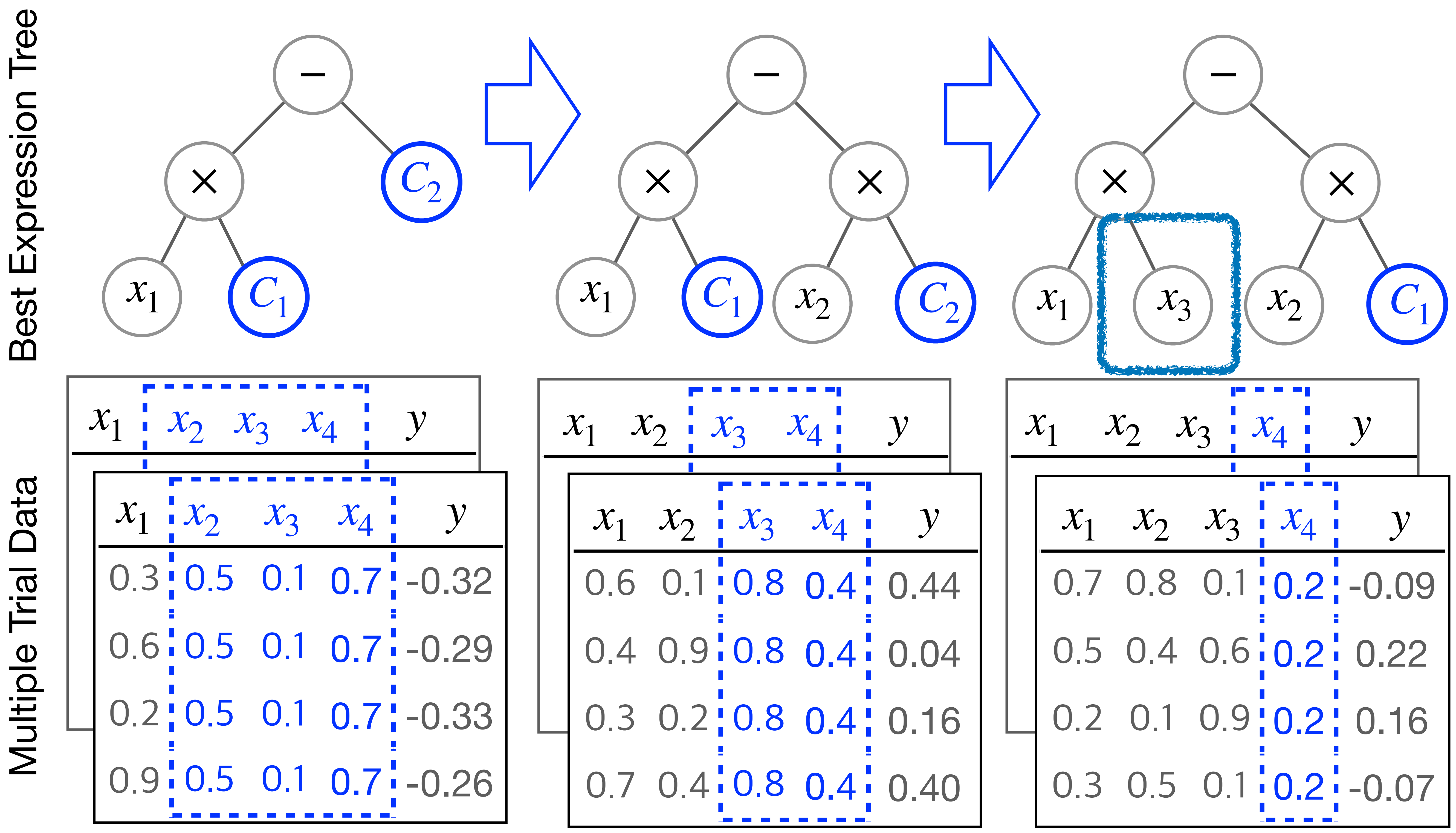
- Assumption: need a data oracle that can return the controlled variables dataset
- We can iteratively reduce the number of controlled variables.



(a) control x_2, x_3, x_4 **(b)** control x_3, x_4 **(c)** control x_4

Control Variable Genetic Programming

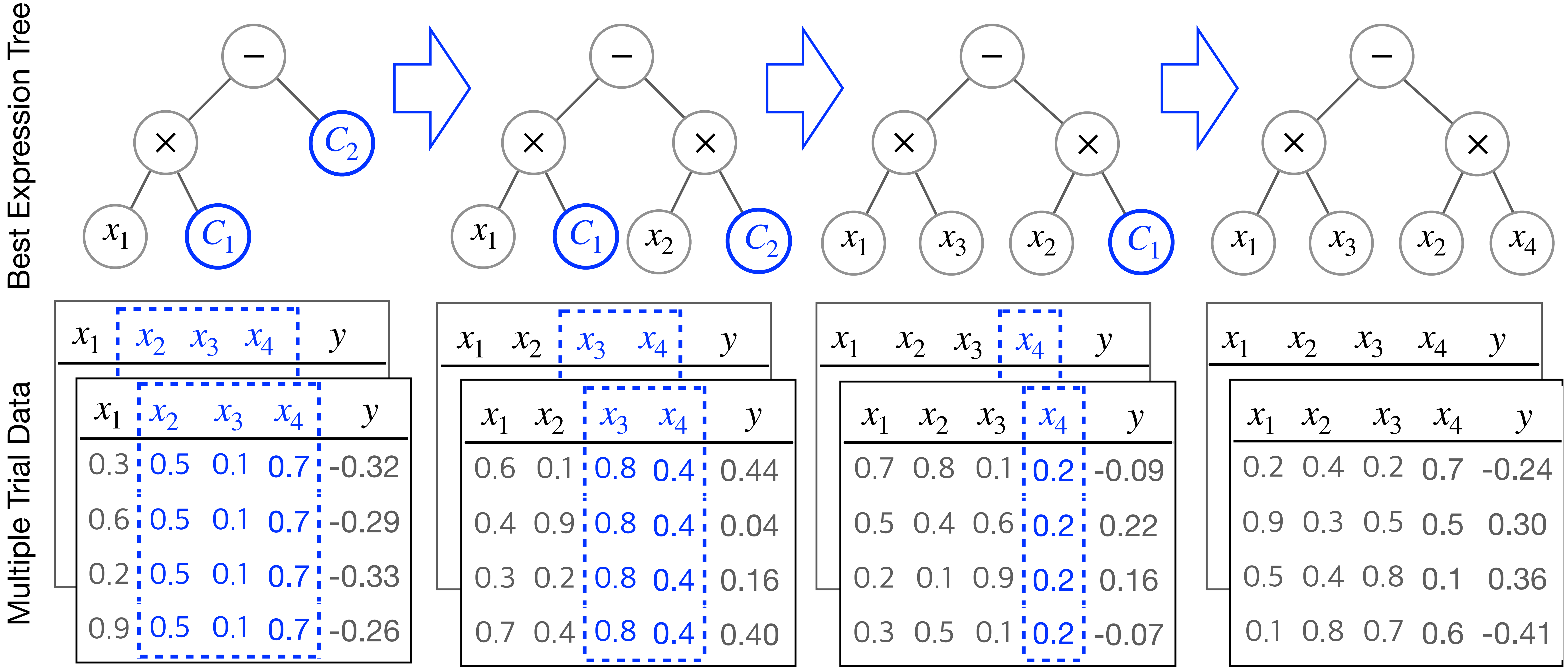
- Assumption: need a data oracle that can return the controlled variables dataset
- We can iteratively reduce the number of controlled variables.



(a) control x_2, x_3, x_4 **(b)** control x_3, x_4 **(c)** control x_4

Control Variable Genetic Programming

- Assumption: need a data oracle that can return the controlled variables dataset
- We can iteratively reduce the number of controlled variables.



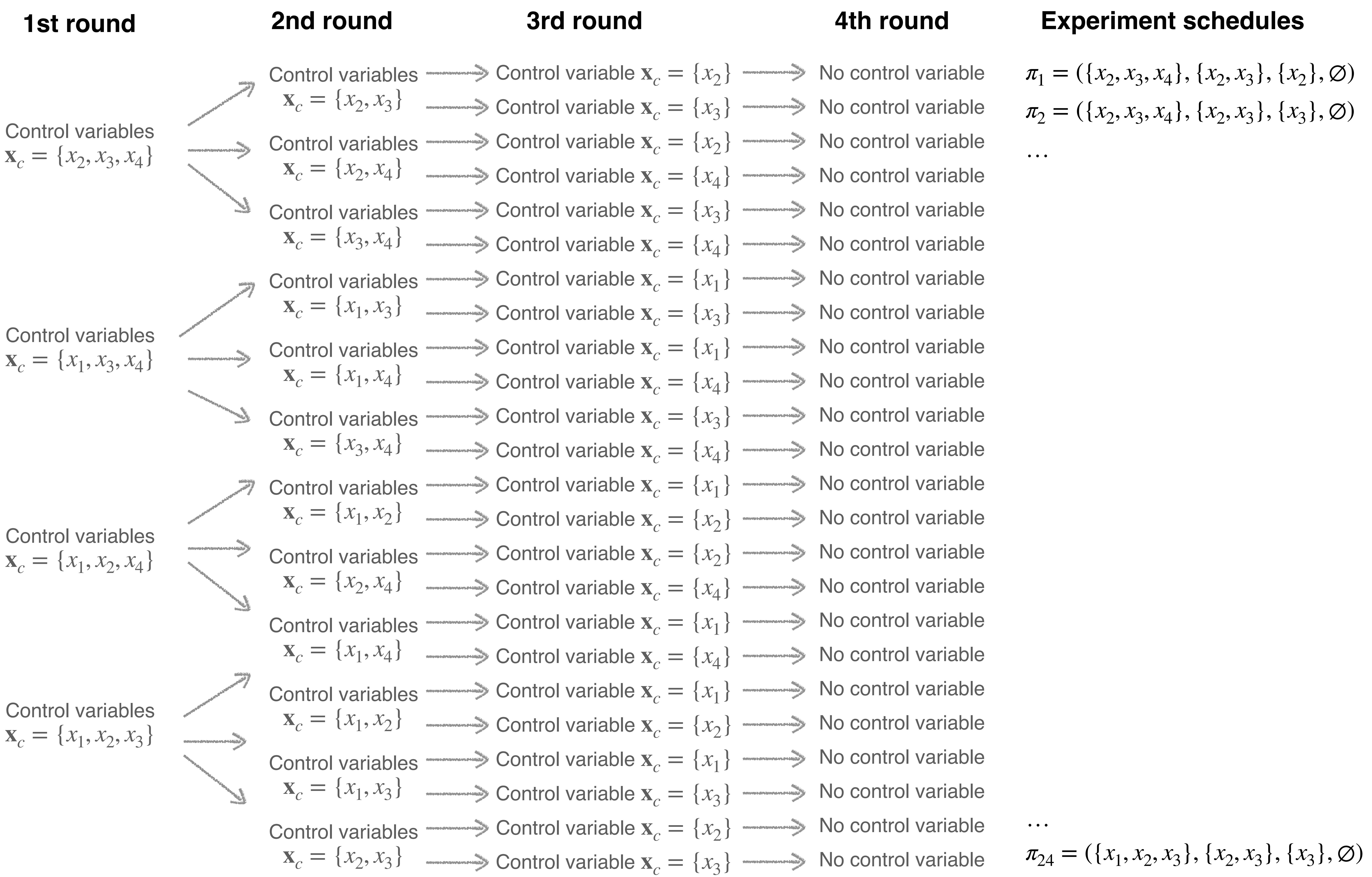
(a) control x_2, x_3, x_4 **(b)** control x_3, x_4 **(c)** control x_4 **(d)** no control

Experiment schedules:

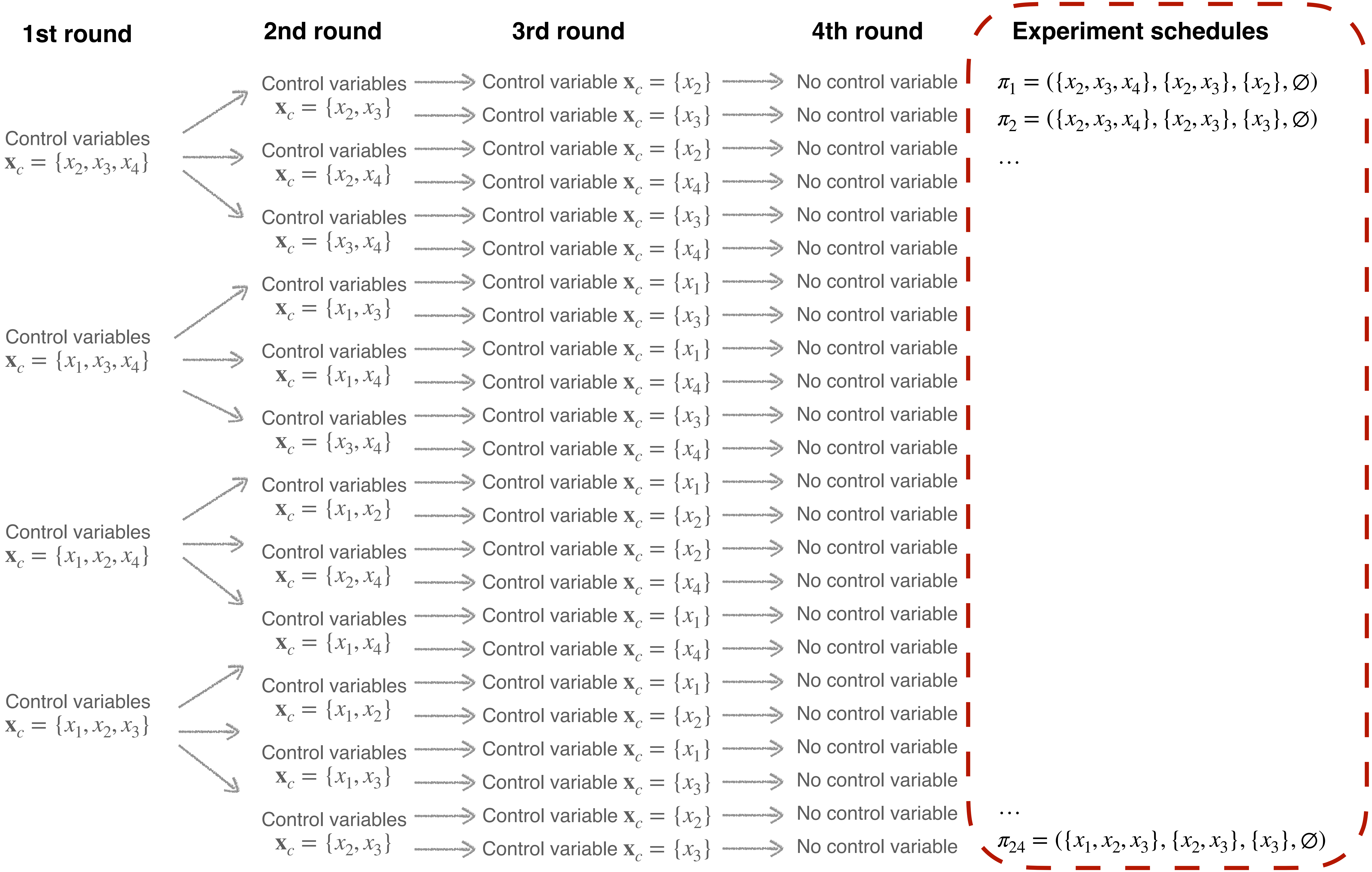
All the possible combinations of controlled variables at every rounds

1st round | **2nd round** | **3rd round** | **4th round**

- Control variables $\mathbf{x}_c = \{x_2, x_3, x_4\}$
  - Control variables $\mathbf{x}_c = \{x_2, x_3\}$
    - Control variable $\mathbf{x}_c = \{x_2\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_3\}$ → No control variable
  - Control variables $\mathbf{x}_c = \{x_2, x_4\}$
    - Control variable $\mathbf{x}_c = \{x_2\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_4\}$ → No control variable
  - Control variables $\mathbf{x}_c = \{x_3, x_4\}$
    - Control variable $\mathbf{x}_c = \{x_3\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_4\}$ → No control variable
- Control variables $\mathbf{x}_c = \{x_1, x_3, x_4\}$
  - Control variables $\mathbf{x}_c = \{x_1, x_3\}$
    - Control variable $\mathbf{x}_c = \{x_1\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_3\}$ → No control variable
  - Control variables $\mathbf{x}_c = \{x_1, x_4\}$
    - Control variable $\mathbf{x}_c = \{x_1\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_4\}$ → No control variable
  - Control variables $\mathbf{x}_c = \{x_3, x_4\}$
    - Control variable $\mathbf{x}_c = \{x_3\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_4\}$ → No control variable
- Control variables $\mathbf{x}_c = \{x_1, x_2, x_4\}$
  - Control variables $\mathbf{x}_c = \{x_1, x_2\}$
    - Control variable $\mathbf{x}_c = \{x_1\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_2\}$ → No control variable
  - Control variables $\mathbf{x}_c = \{x_2, x_4\}$
    - Control variable $\mathbf{x}_c = \{x_2\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_4\}$ → No control variable
  - Control variables $\mathbf{x}_c = \{x_1, x_4\}$
    - Control variable $\mathbf{x}_c = \{x_1\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_4\}$ → No control variable
- Control variables $\mathbf{x}_c = \{x_1, x_2, x_3\}$
  - Control variables $\mathbf{x}_c = \{x_1, x_2\}$
    - Control variable $\mathbf{x}_c = \{x_1\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_2\}$ → No control variable
  - Control variables $\mathbf{x}_c = \{x_1, x_3\}$
    - Control variable $\mathbf{x}_c = \{x_1\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_3\}$ → No control variable
  - Control variables $\mathbf{x}_c = \{x_2, x_3\}$
    - Control variable $\mathbf{x}_c = \{x_2\}$ → No control variable
    - Control variable $\mathbf{x}_c = \{x_3\}$ → No control variable



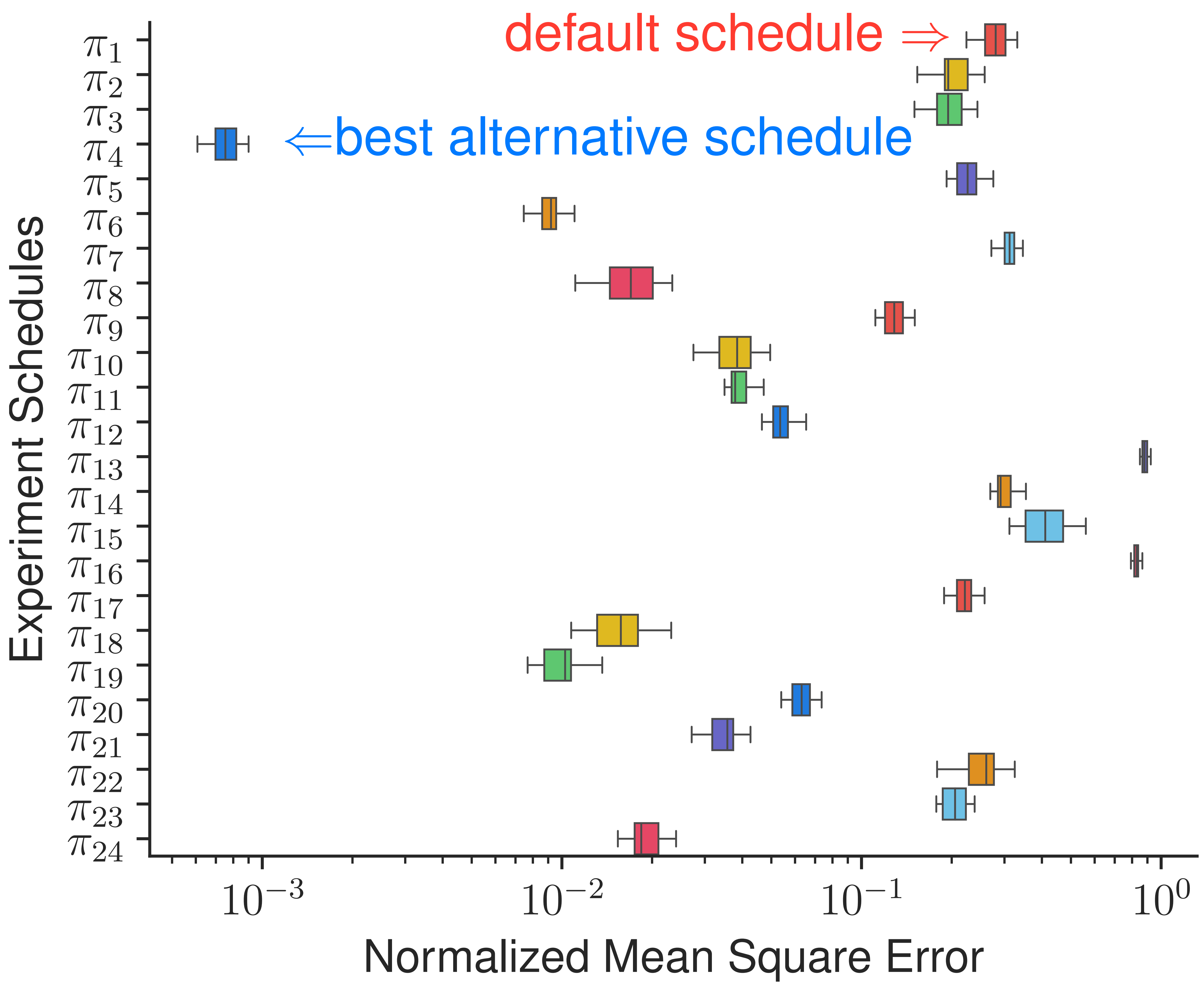
1st round
2nd round
3rd round
4th round
Experiment schedules
Control variables $\mathbf{x}_c = \{x_2, x_3, x_4\}$
Control variables $\mathbf{x}_c = \{x_2, x_3\}$
Control variable $\mathbf{x}_c = \{x_2\}$
No control variable
Control variable $\mathbf{x}_c = \{x_3\}$
No control variable
Control variables $\mathbf{x}_c = \{x_2, x_4\}$
Control variable $\mathbf{x}_c = \{x_2\}$
No control variable
Control variable $\mathbf{x}_c = \{x_4\}$
No control variable
Control variables $\mathbf{x}_c = \{x_3, x_4\}$
Control variable $\mathbf{x}_c = \{x_3\}$
No control variable
Control variable $\mathbf{x}_c = \{x_4\}$
No control variable
Control variables $\mathbf{x}_c = \{x_1, x_3, x_4\}$
Control variables $\mathbf{x}_c = \{x_1, x_3\}$
Control variable $\mathbf{x}_c = \{x_1\}$
No control variable
Control variable $\mathbf{x}_c = \{x_3\}$
No control variable
Control variables $\mathbf{x}_c = \{x_1, x_4\}$
Control variable $\mathbf{x}_c = \{x_1\}$
No control variable
Control variable $\mathbf{x}_c = \{x_4\}$
No control variable
Control variables $\mathbf{x}_c = \{x_3, x_4\}$
Control variable $\mathbf{x}_c = \{x_3\}$
No control variable
Control variable $\mathbf{x}_c = \{x_4\}$
No control variable
Control variables $\mathbf{x}_c = \{x_1, x_2, x_4\}$
Control variables $\mathbf{x}_c = \{x_1, x_2\}$
Control variable $\mathbf{x}_c = \{x_1\}$
No control variable
Control variable $\mathbf{x}_c = \{x_2\}$
No control variable
Control variables $\mathbf{x}_c = \{x_2, x_4\}$
Control variable $\mathbf{x}_c = \{x_2\}$
No control variable
Control variable $\mathbf{x}_c = \{x_4\}$
No control variable
Control variables $\mathbf{x}_c = \{x_1, x_4\}$
Control variable $\mathbf{x}_c = \{x_1\}$
No control variable
Control variable $\mathbf{x}_c = \{x_4\}$
No control variable
Control variables $\mathbf{x}_c = \{x_1, x_2, x_3\}$
Control variables $\mathbf{x}_c = \{x_1, x_2\}$
Control variable $\mathbf{x}_c = \{x_1\}$
No control variable
Control variable $\mathbf{x}_c = \{x_2\}$
No control variable
Control variables $\mathbf{x}_c = \{x_1, x_3\}$
Control variable $\mathbf{x}_c = \{x_1\}$
No control variable
Control variable $\mathbf{x}_c = \{x_3\}$
No control variable
Control variables $\mathbf{x}_c = \{x_2, x_3\}$
Control variable $\mathbf{x}_c = \{x_2\}$
No control variable
Control variable $\mathbf{x}_c = \{x_3\}$
No control variable
$\pi_1 = (\{x_2, x_3, x_4\}, \{x_2, x_3\}, \{x_2\}, \varnothing)$
$\pi_2 = (\{x_2, x_3, x_4\}, \{x_2, x_3\}, \{x_3\}, \varnothing)$
$\cdots$
$\cdots$
$\pi_{24} = (\{x_1, x_2, x_3\}, \{x_2, x_3\}, \{x_3\}, \varnothing)$

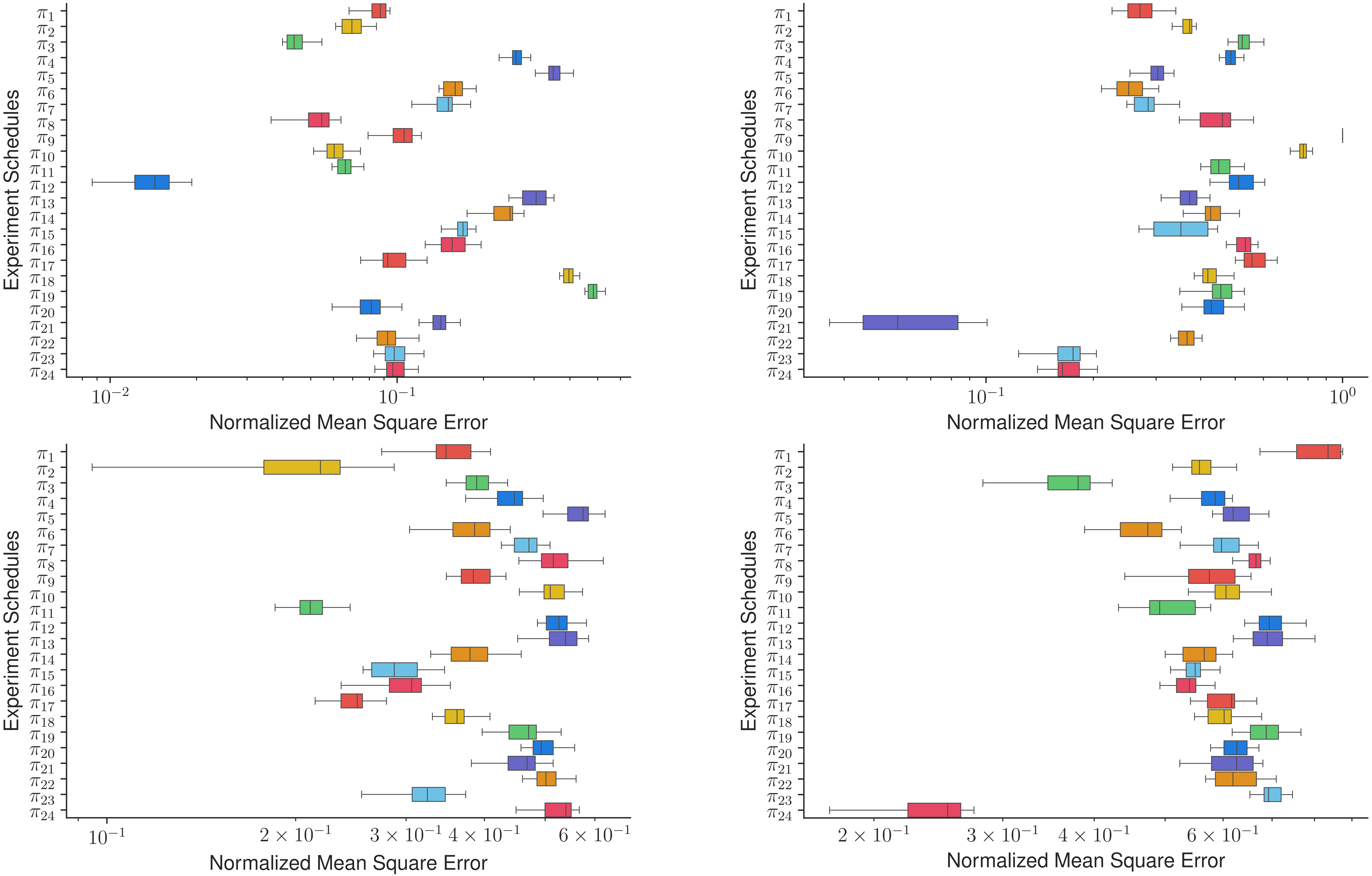


1st round
2nd round
3rd round
4th round
Experiment schedules
Control variables $\mathbf{x}_c = \{x_2, x_3, x_4\}$
Control variables $\mathbf{x}_c = \{x_1, x_3, x_4\}$
Control variables $\mathbf{x}_c = \{x_1, x_2, x_4\}$
Control variables $\mathbf{x}_c = \{x_1, x_2, x_3\}$
Control variables $\mathbf{x}_c = \{x_2, x_3\}$
Control variables $\mathbf{x}_c = \{x_2, x_4\}$
Control variables $\mathbf{x}_c = \{x_3, x_4\}$
Control variables $\mathbf{x}_c = \{x_1, x_3\}$
Control variables $\mathbf{x}_c = \{x_1, x_4\}$
Control variables $\mathbf{x}_c = \{x_3, x_4\}$
Control variables $\mathbf{x}_c = \{x_1, x_2\}$
Control variables $\mathbf{x}_c = \{x_2, x_4\}$
Control variables $\mathbf{x}_c = \{x_1, x_4\}$
Control variables $\mathbf{x}_c = \{x_1, x_2\}$
Control variables $\mathbf{x}_c = \{x_1, x_3\}$
Control variables $\mathbf{x}_c = \{x_2, x_3\}$
Control variable $\mathbf{x}_c = \{x_2\}$
Control variable $\mathbf{x}_c = \{x_3\}$
Control variable $\mathbf{x}_c = \{x_2\}$
Control variable $\mathbf{x}_c = \{x_4\}$
Control variable $\mathbf{x}_c = \{x_3\}$
Control variable $\mathbf{x}_c = \{x_4\}$
Control variable $\mathbf{x}_c = \{x_1\}$
Control variable $\mathbf{x}_c = \{x_3\}$
Control variable $\mathbf{x}_c = \{x_1\}$
Control variable $\mathbf{x}_c = \{x_4\}$
Control variable $\mathbf{x}_c = \{x_3\}$
Control variable $\mathbf{x}_c = \{x_4\}$
Control variable $\mathbf{x}_c = \{x_1\}$
Control variable $\mathbf{x}_c = \{x_2\}$
Control variable $\mathbf{x}_c = \{x_2\}$
Control variable $\mathbf{x}_c = \{x_4\}$
Control variable $\mathbf{x}_c = \{x_1\}$
Control variable $\mathbf{x}_c = \{x_4\}$
Control variable $\mathbf{x}_c = \{x_1\}$
Control variable $\mathbf{x}_c = \{x_2\}$
Control variable $\mathbf{x}_c = \{x_1\}$
Control variable $\mathbf{x}_c = \{x_3\}$
Control variable $\mathbf{x}_c = \{x_2\}$
Control variable $\mathbf{x}_c = \{x_3\}$
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
No control variable
$\pi_1 = (\{x_2, x_3, x_4\}, \{x_2, x_3\}, \{x_2\}, \varnothing)$
$\pi_2 = (\{x_2, x_3, x_4\}, \{x_2, x_3\}, \{x_3\}, \varnothing)$
$\dots$
$\dots$
$\pi_{24} = (\{x_1, x_2, x_3\}, \{x_2, x_3\}, \{x_3\}, \varnothing)$

CVGP is sensitive to experiment schedules.

- There exists a better experiment schedule (i.e., π_4) among all schedules than the default one (i.e., π_1), in terms of NMSE metric.

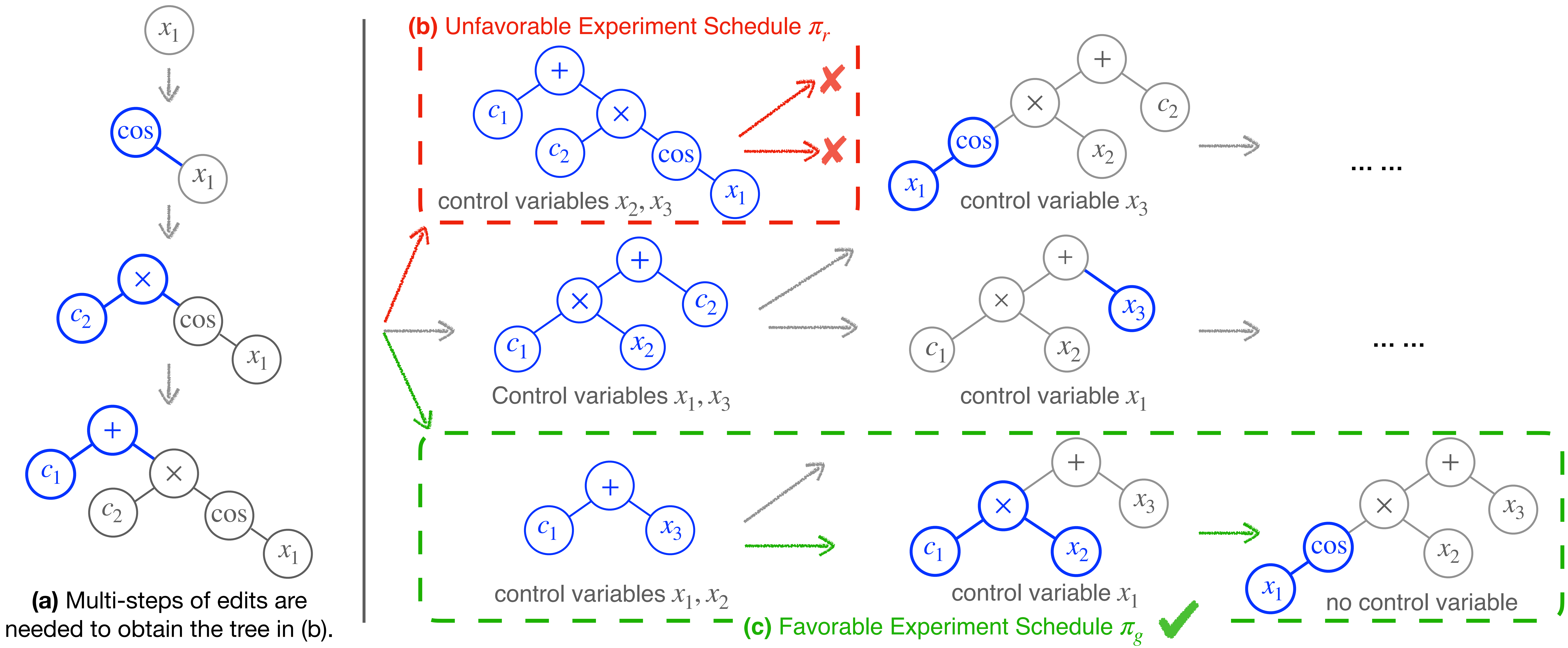




Experiment Schedules
π_1 π_2 π_3 π_4 π_5 π_6 π_7 π_8 π_9 π_{10} π_{11} π_{12} π_{13} π_{14} π_{15} π_{16} π_{17} π_{18} π_{19} π_{20} π_{21} π_{22} π_{23} π_{24}
10^{-2} 10^{-1}
Normalized Mean Square Error
Experiment Schedules
10^{-1} 10^{0}
Normalized Mean Square Error
Experiment Schedules
10^{-1} 2×10^{-1} 3×10^{-1} 4×10^{-1} 6×10^{-1}
Normalized Mean Square Error
Experiment Schedules
2×10^{-1} 3×10^{-1} 4×10^{-1} 6×10^{-1}
Normalized Mean Square Error

Our idea: Racing CVGP

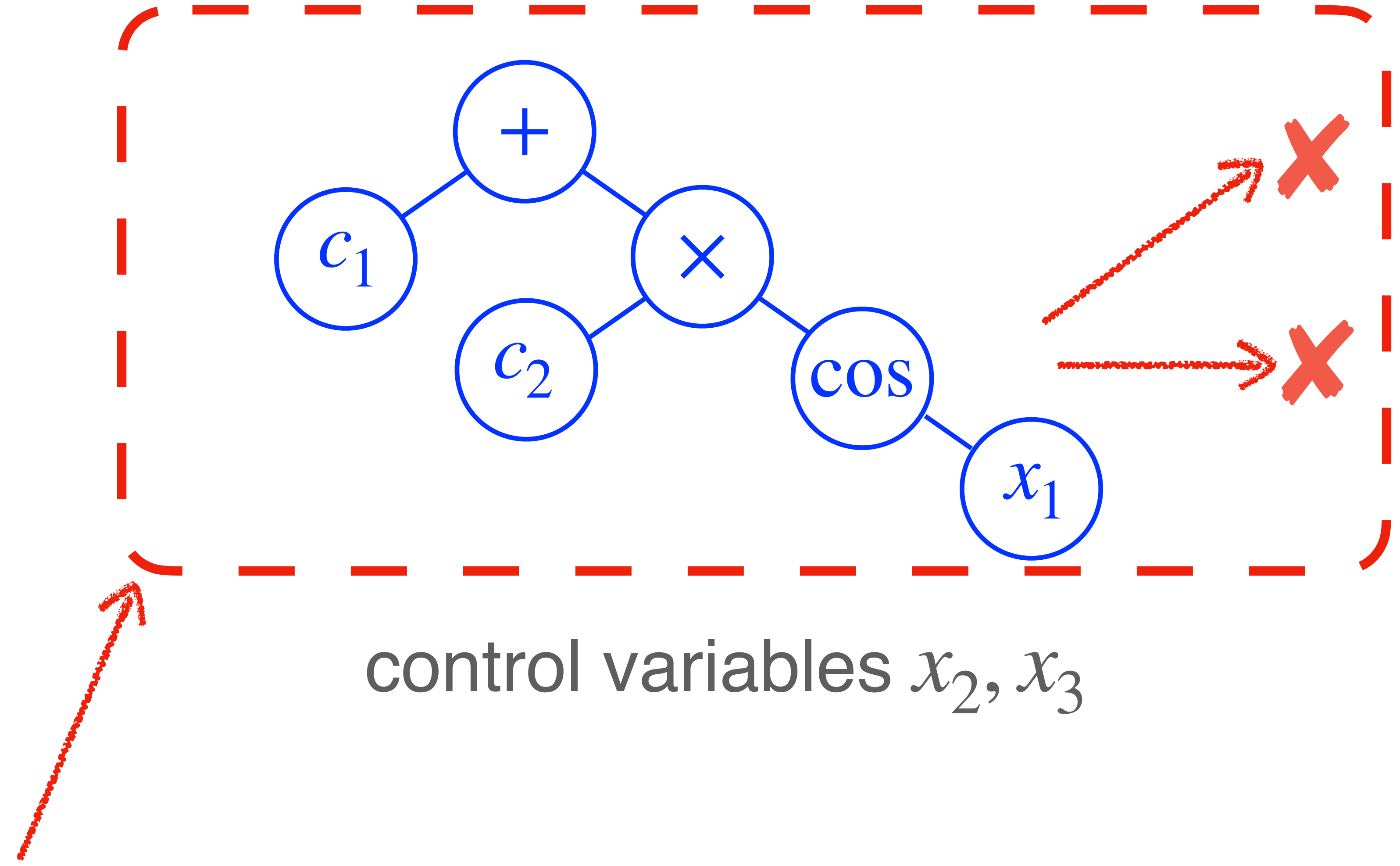
(1) maintaining ***multiple*** experiment schedules rather than one.
(2) allowing **promising** experiment schedules to **survive** while letting **unfavorable** schedules **early stop**.



Our idea: Racing CVGP

We want to discover the expression $\phi = \cos(x_1) \times x_2 + x_3$.

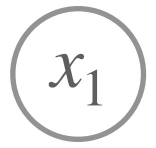
(b) Unfavorable Experiment Schedule π_r



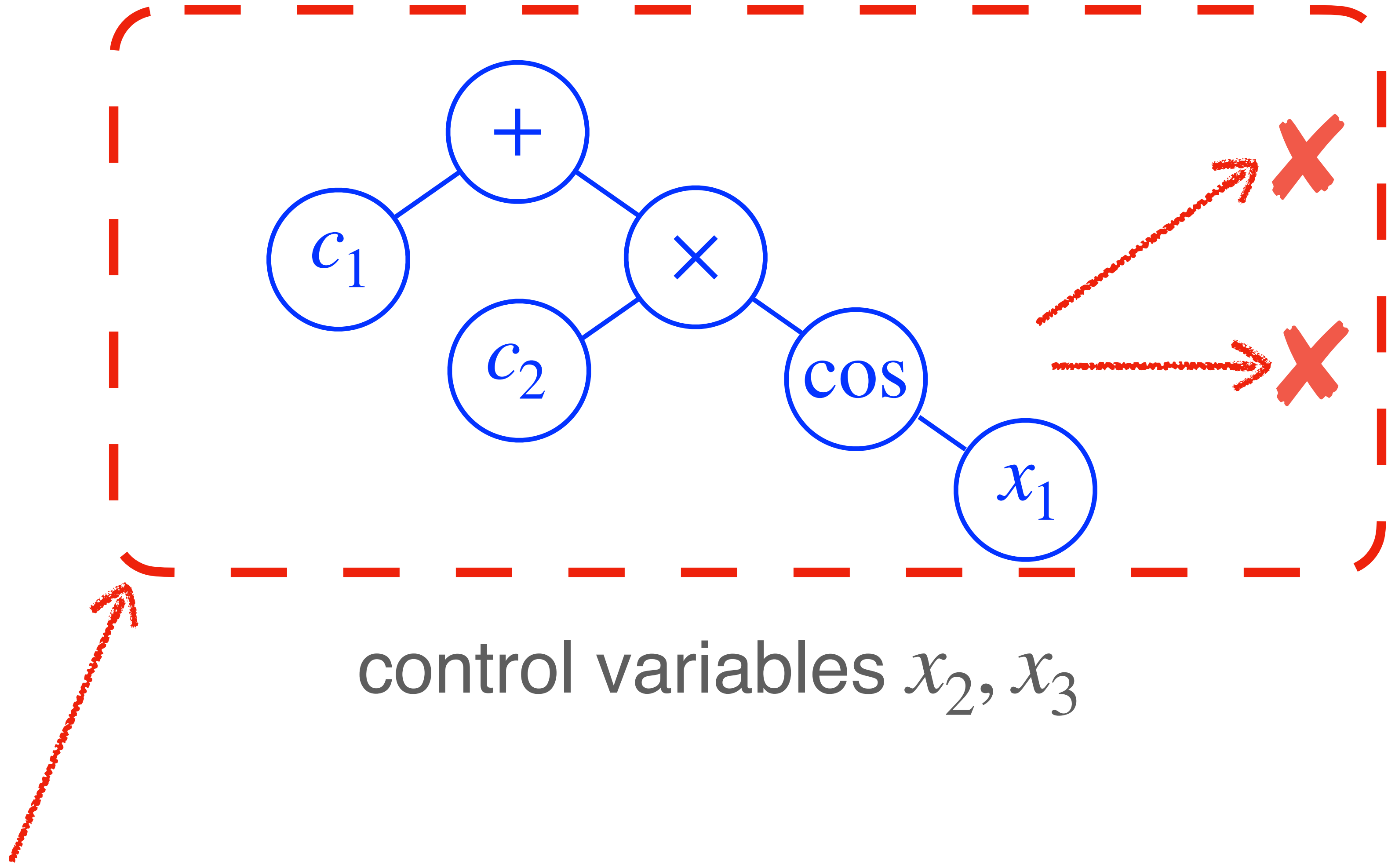
control variables x_2, x_3

Our idea: Racing CVGP

We want to discover the expression $\phi = \cos(x_1) \times x_2 + x_3$.



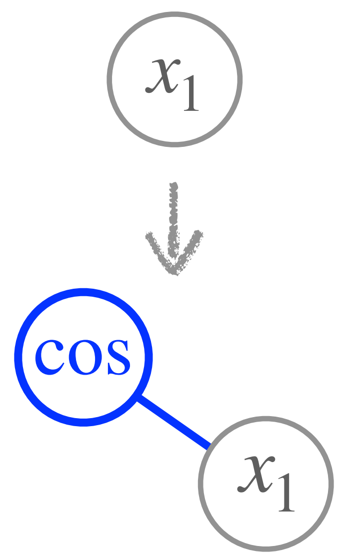
(b) Unfavorable Experiment Schedule π_r

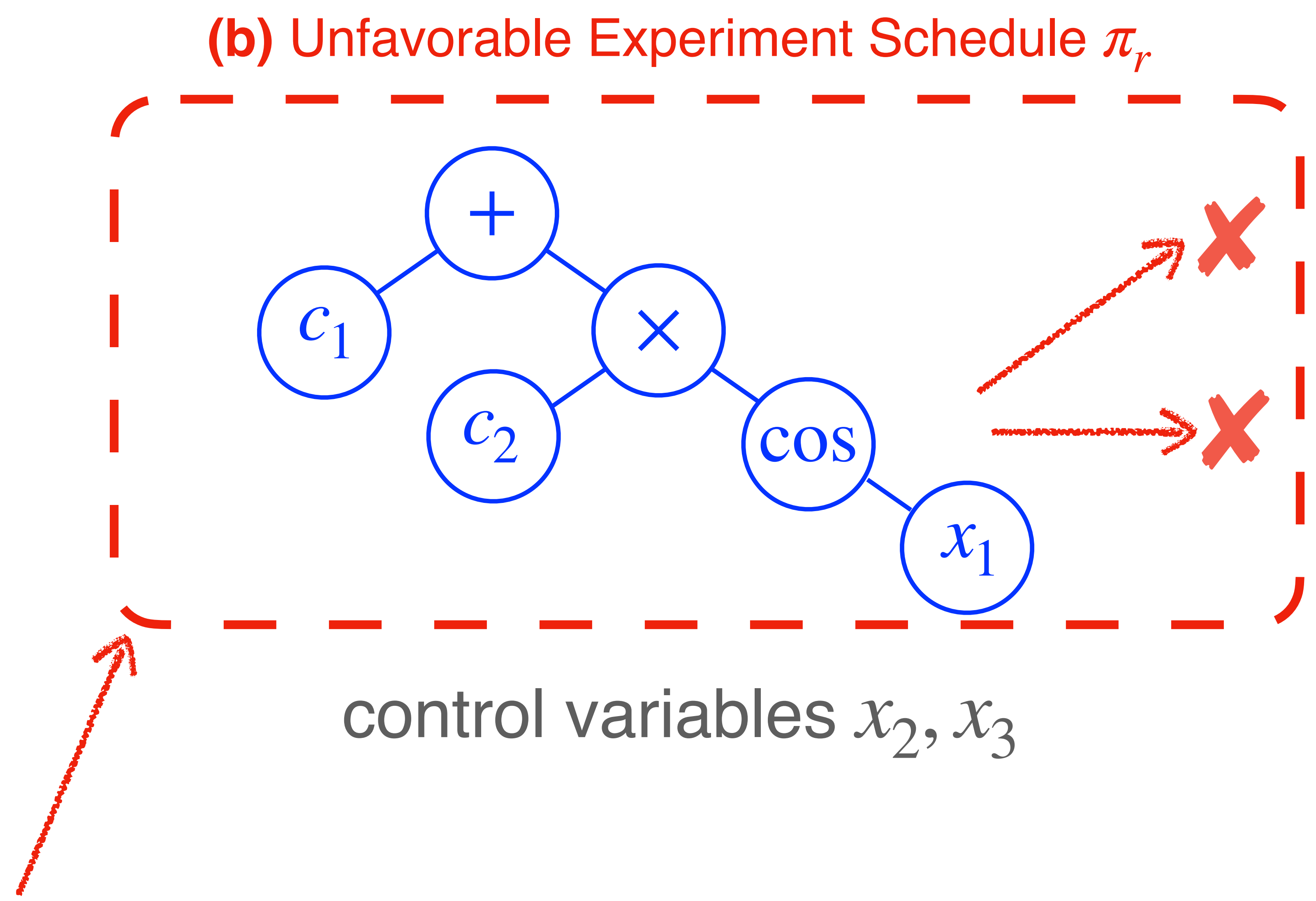


control variables x_2, x_3

Our idea: Racing CVGP

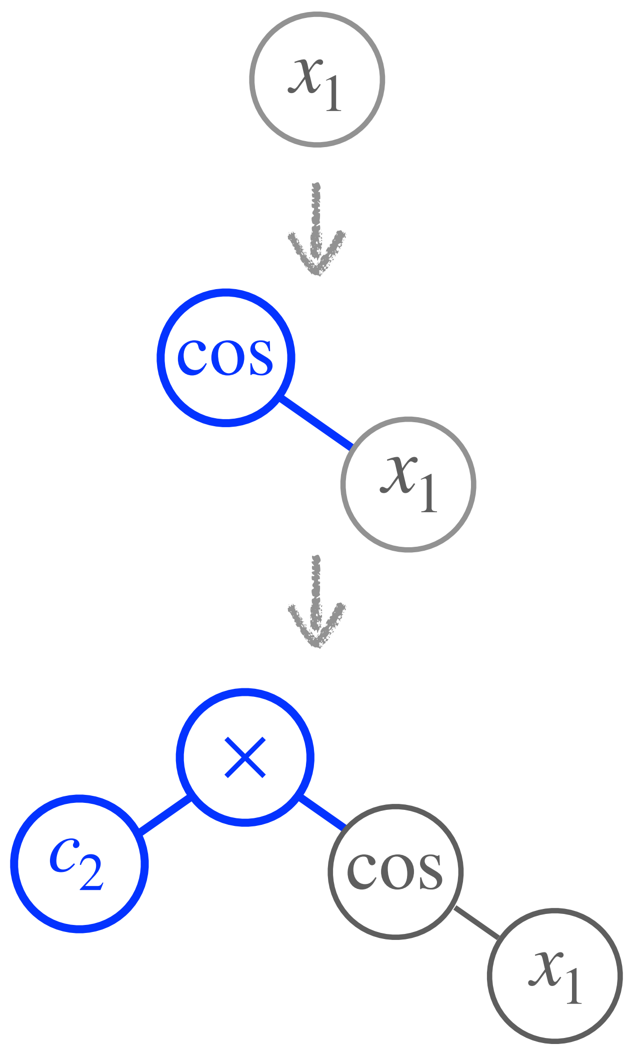
We want to discover the expression $\phi = \cos(x_1) \times x_2 + x_3$.

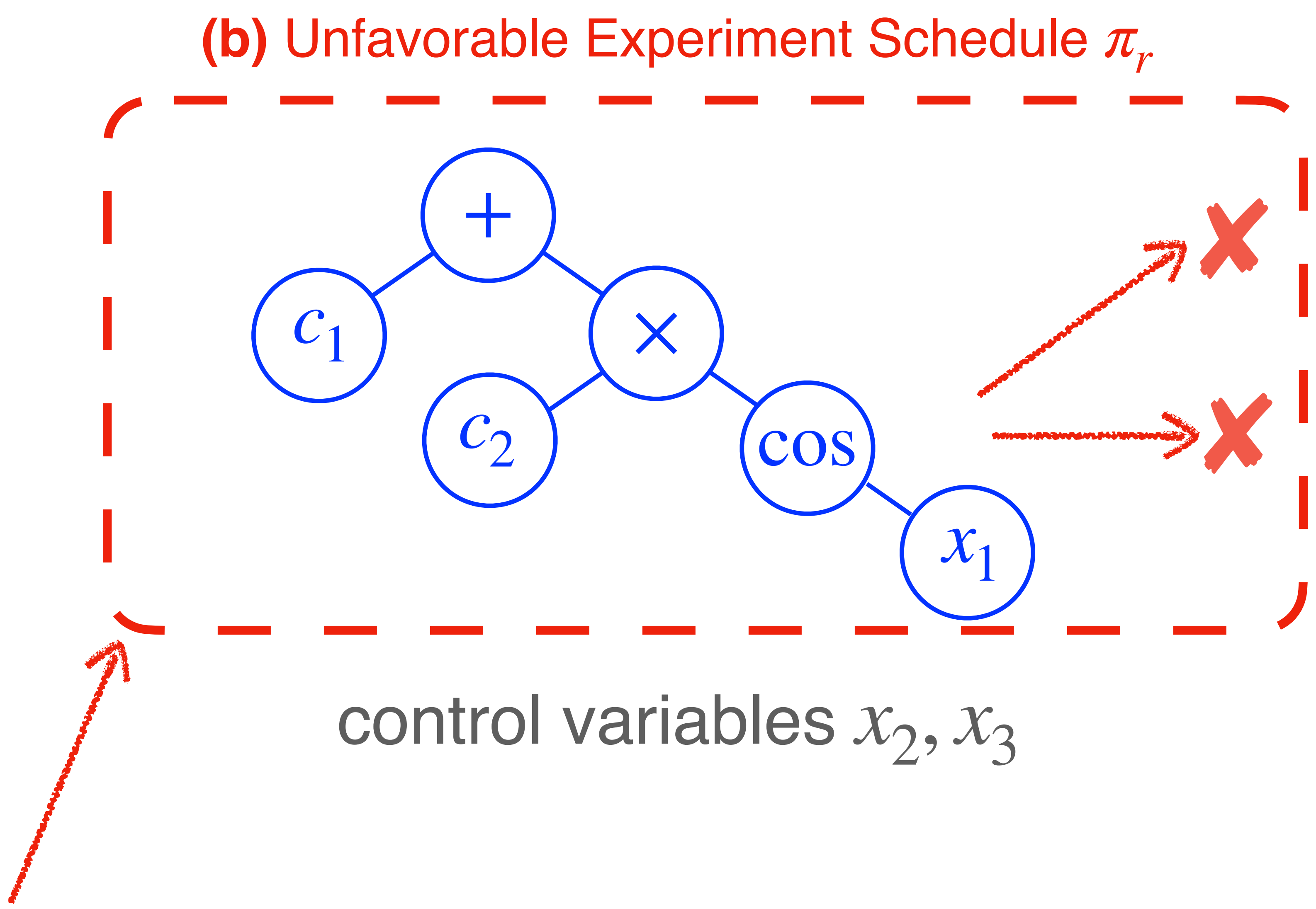




Our idea: Racing CVGP

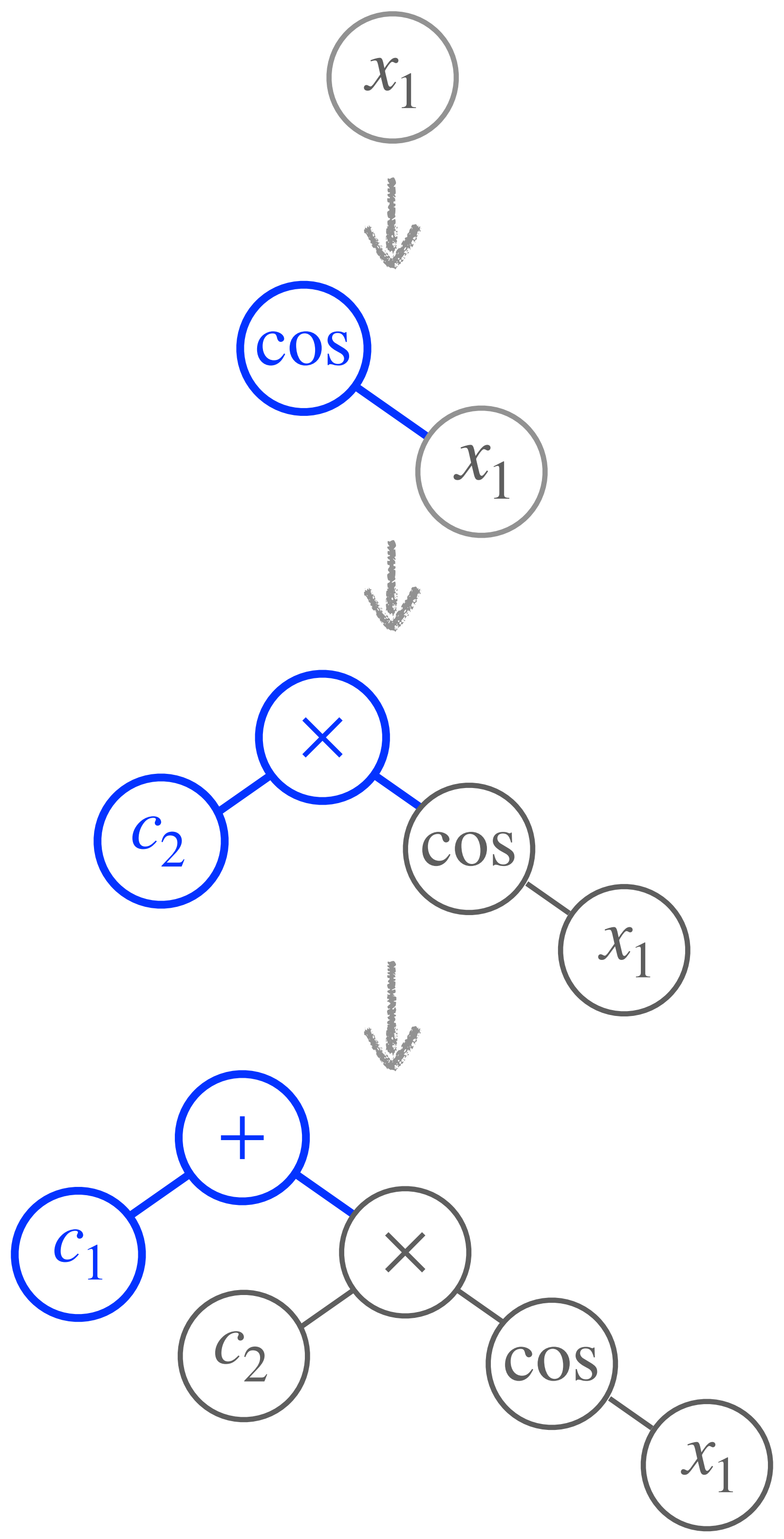
We want to discover the expression $\phi = \cos(x_1) \times x_2 + x_3$.

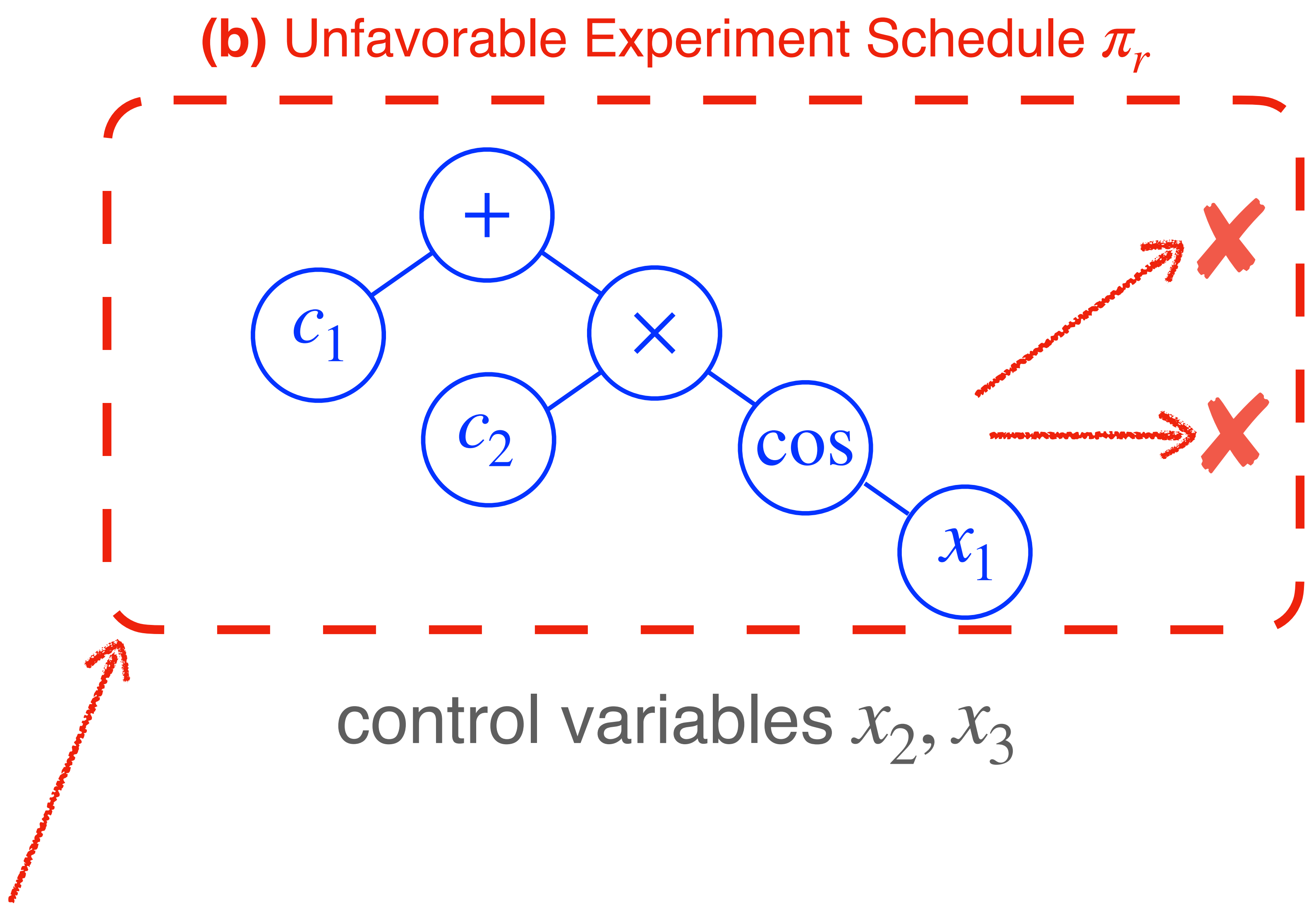




Our idea: Racing CVGP

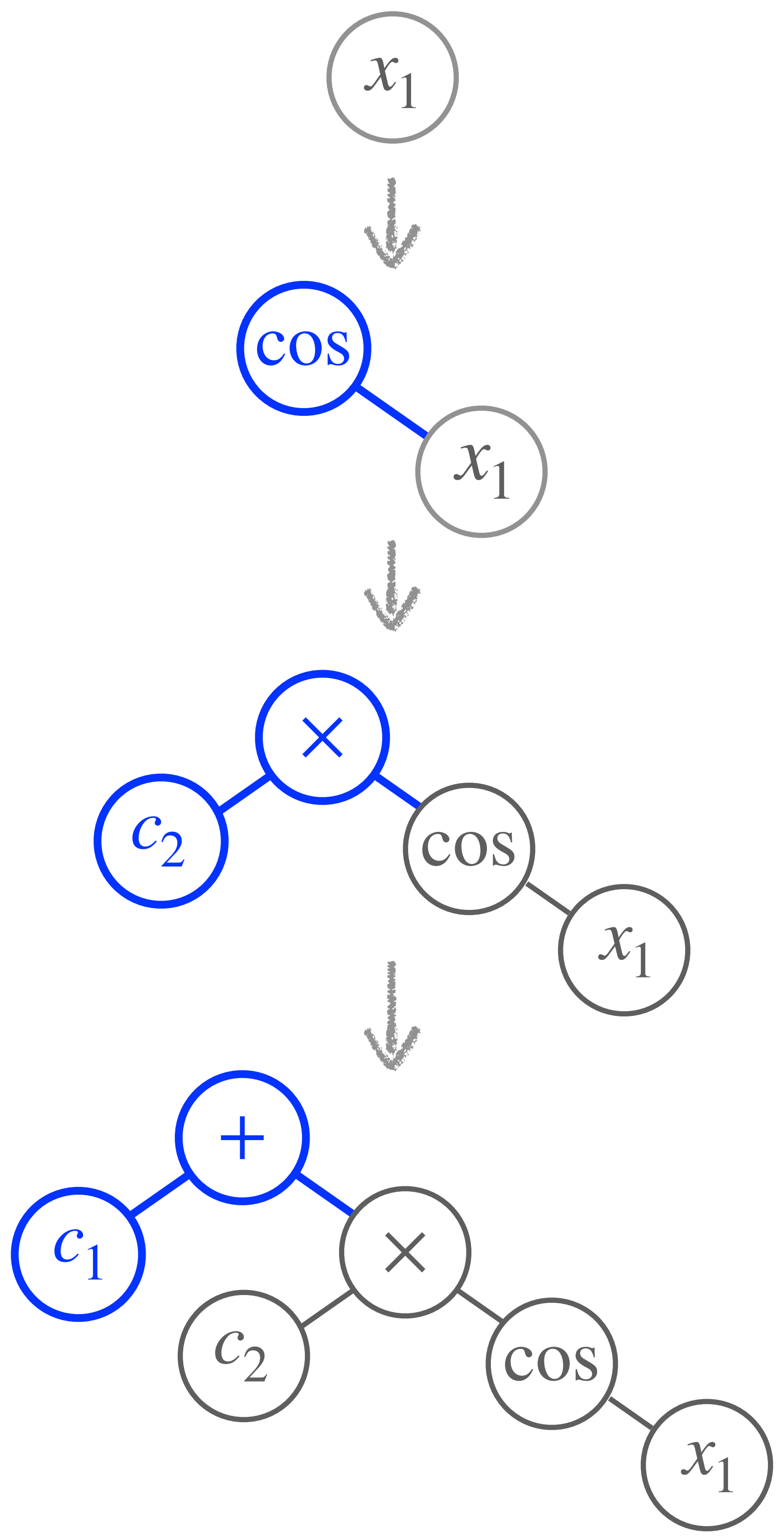
We want to discover the expression $\phi = \cos(x_1) \times x_2 + x_3$.

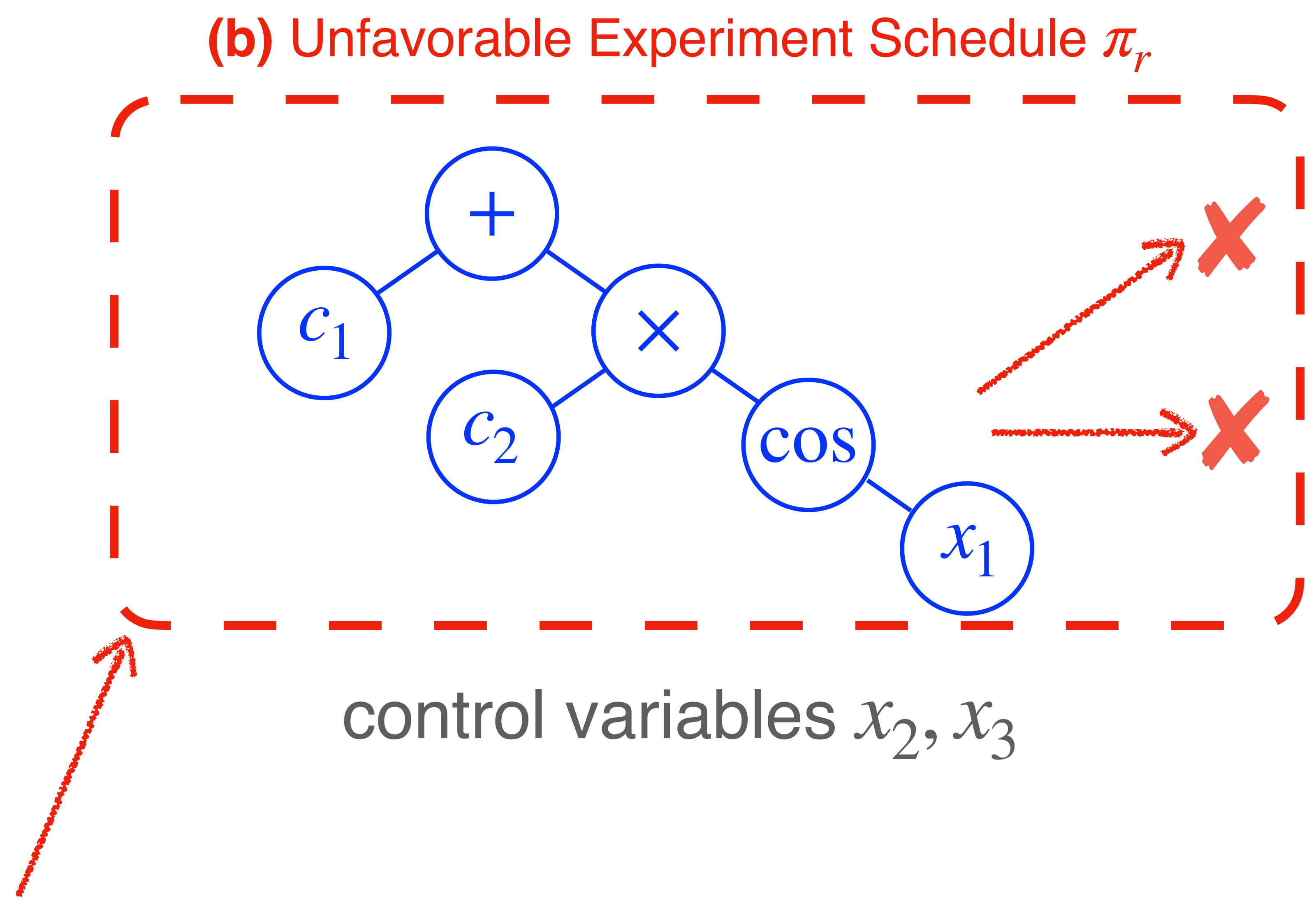




Our idea: Racing CVGP

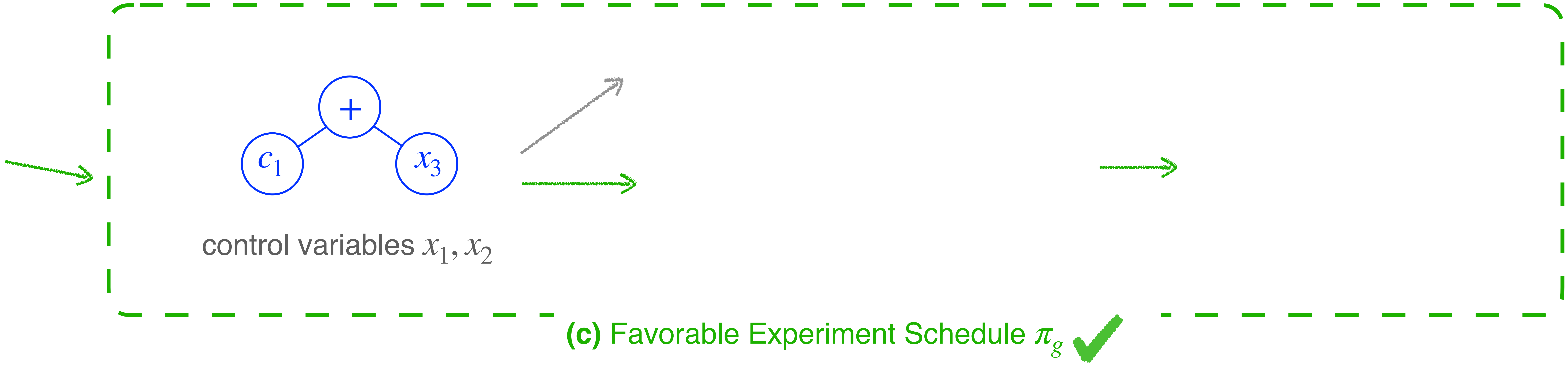
We want to discover the expression $\phi = \cos(x_1) \times x_2 + x_3$.





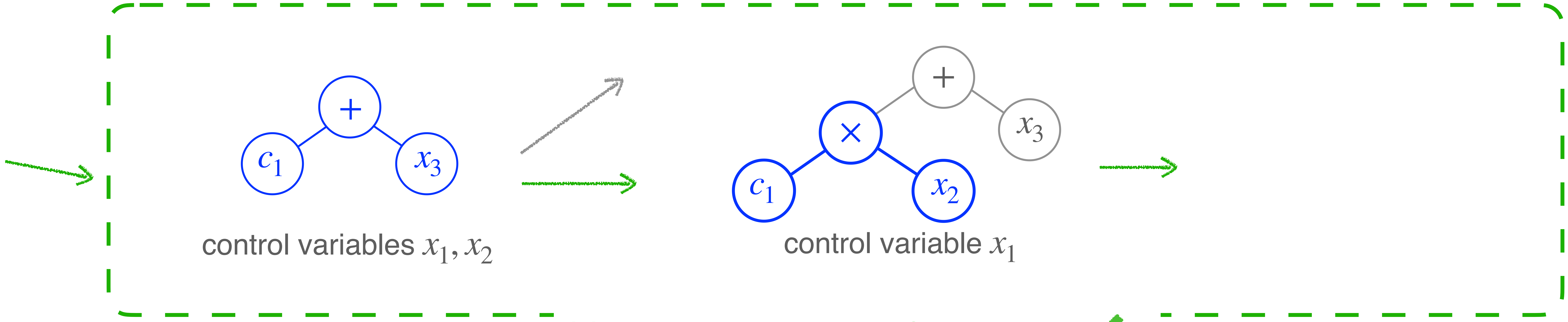
Multi-steps of edits are needed by GP to obtain the expression tree

Our idea: Racing CVGP



- it is relatively easy to discover following the green schedule
- every change in the expression tree is reasonable

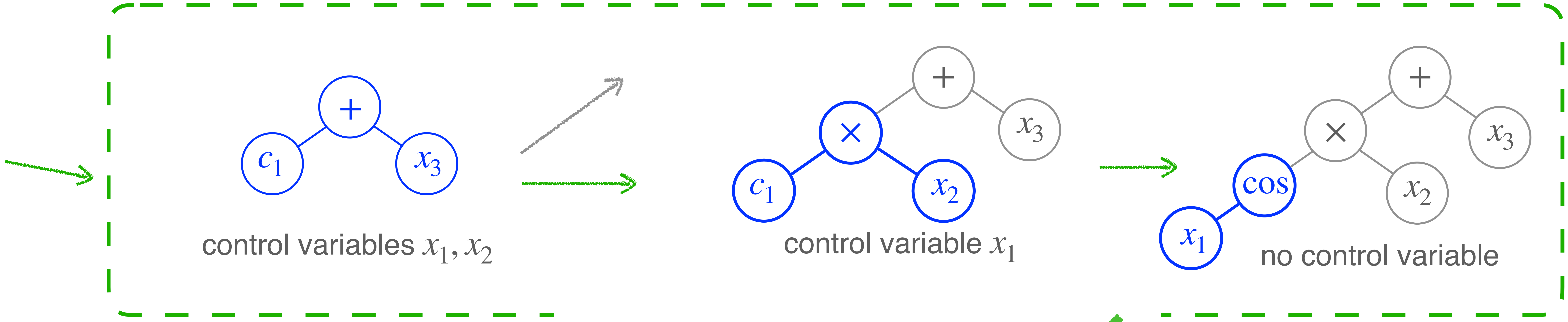
Our idea: Racing CVGP



(c) Favorable Experiment Schedule π_g

- it is relatively easy to discover following the green schedule
- every change in the expression tree is reasonable

Our idea: Racing CVGP

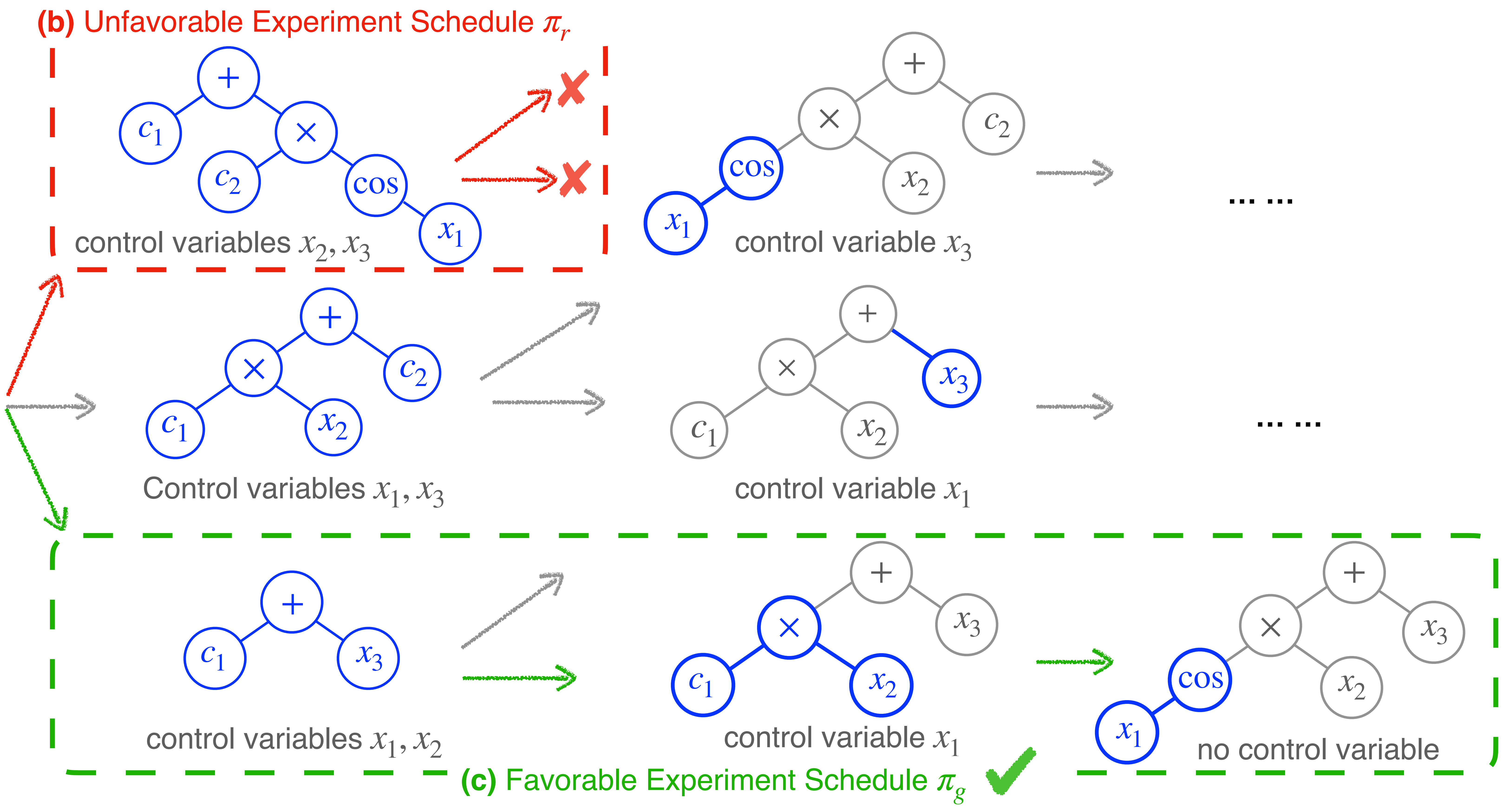


(c) Favorable Experiment Schedule π_g

- it is relatively easy to discover following the green schedule
- every change in the expression tree is reasonable

Our idea: Racing CVGP

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



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.

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.

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.
- Another implicit factor is the number of constants in each expression (see experiments).

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

Experiments Analysis

Goodness-of-fit Benchmark

NMSE value (normalized mean squared error)

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

Goodness-of-fit Benchmark

NMSE value (normalized mean squared error)

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

Goodness-of-fit Benchmark

NMSE value (normalized mean squared error)

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

Running Time comparison

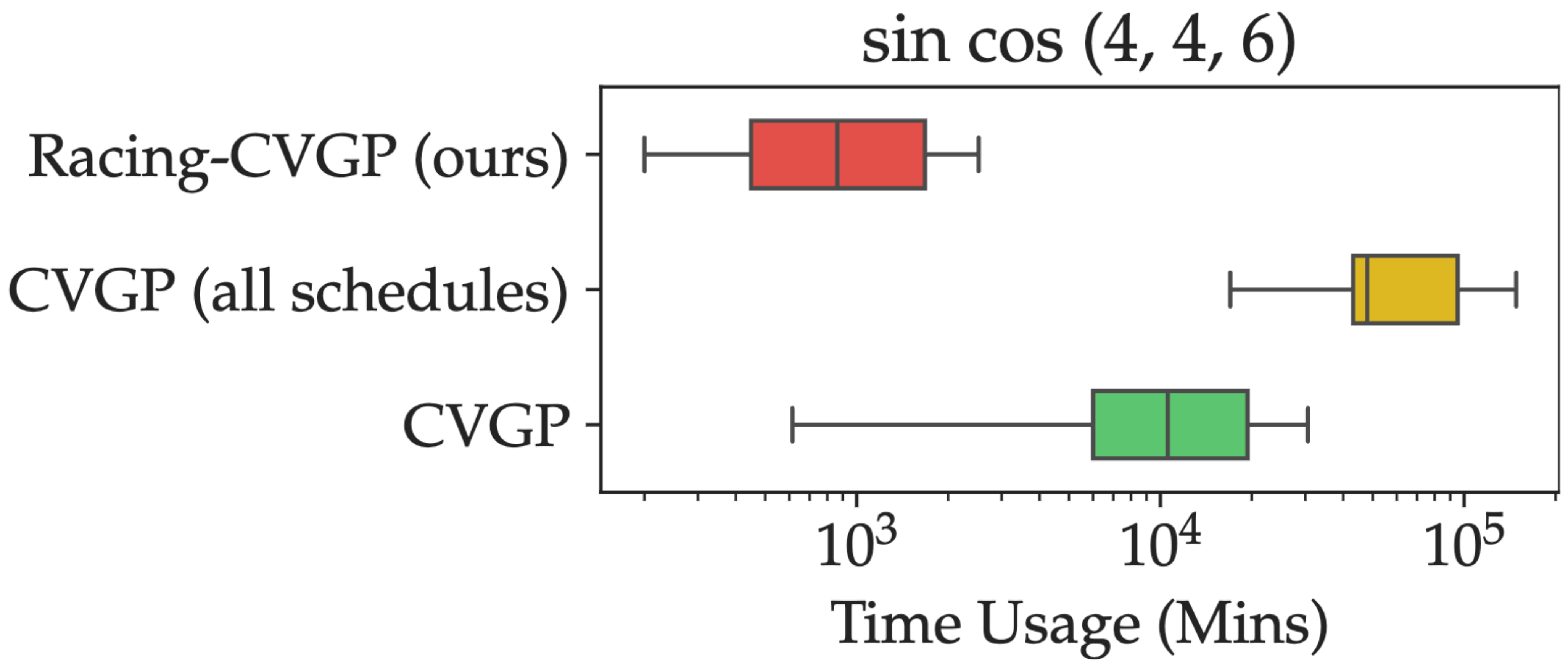


Figure 6: On a selected Trigonometric dataset, quartiles of the total running time of Racing-CVGP, CVGP, and CVGP with all the experiment schedules. Our Racing-CVGP saves a great portion of time compared with CVGP with all the schedules for expressions with $n = 4$ variables.

Running Time comparison

Our Racing-CVGP saves times.

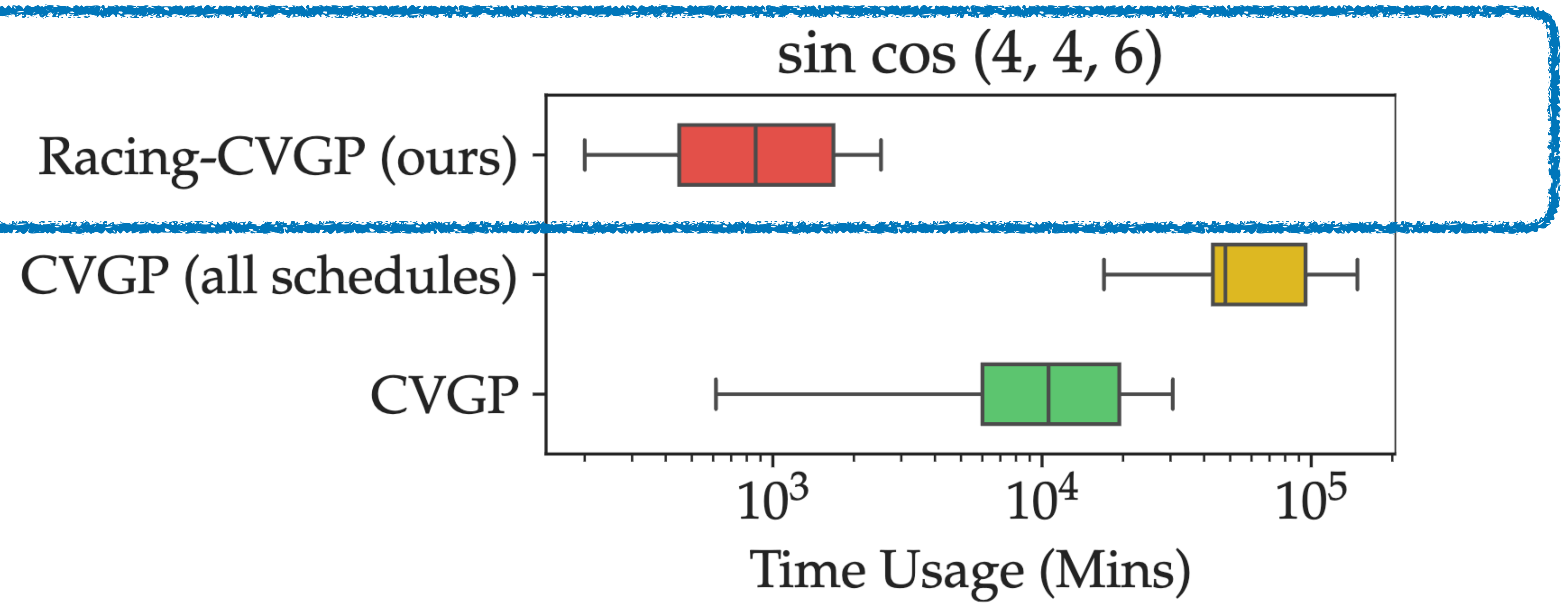


Figure 6: On a selected Trigonometric dataset, quartiles of the total running time of Racing-CVGP, CVGP, and CVGP with all the experiment schedules. Our Racing-CVGP saves a great portion of time compared with CVGP with all the schedules for expressions with $n = 4$ variables.

Conclusion

- We propose Racing Control Variable Genetic Programming to **accelerate** the discovery process of symbolic regression.
- Our idea is to maintain multiple schedules and early stop those unfavorable schedules.
- In experiments, we find racing-CVGP scales up to expressions with 8 variables which is not possible for baselines, including GP, CVGP and GPmeld.

Q & A

https://bitbucket.org/xlnxyx/racing_cvgp