Reverse-Engineering Stochastic Models

Mufaro Machaya and Ian Matsunaga

June 22, 2025

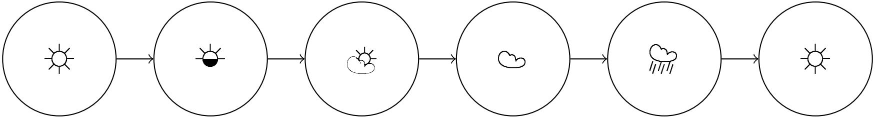
Stochastic

Coins: Heads, Tails, Tails, Heads, Tails, ...

Dice: 0 0 1 0 3 2 5 7 2 5 3 6 7 ...

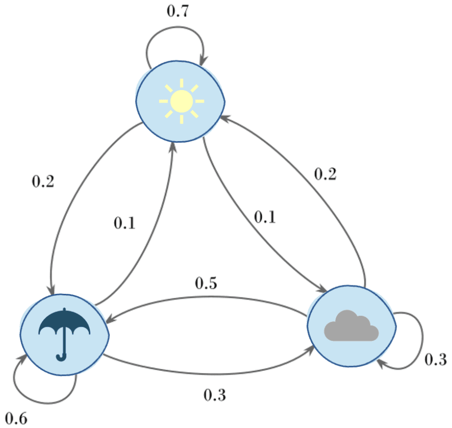
- Conventional modeling techniques are ineffective with stochastic (random) patterns

Weather vs. Time



How could we model this?

Markov Models



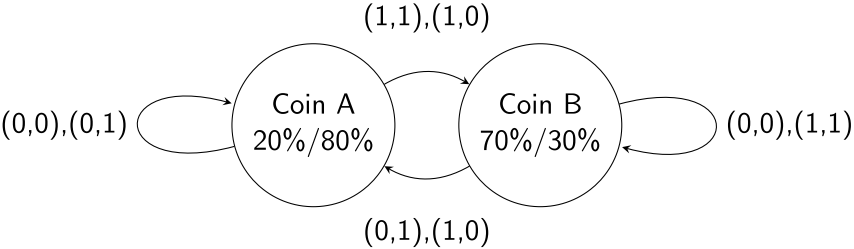
Source: https://www.statology.org/markov-chains-demystified-from-weather-predictions-to-googles-pagerank/.

Core Question and Purpose

How complicated can we make a Markov model before we cannot accurately reverse-engineer it/predict its future behavior?

- Many heavily researched strategies of reverse engineering stochastic models
- Not as much insight into how memory relates the effectiveness of reverse-engineering

Markov Model/Coin Systems



Complexity and Variance

Complexity

$$c = nvm^2 \tag{1}$$

Variance (v)

$$v = \sum_{i=1}^{n} \sum_{j=1}^{n} |p_{i,j} - \mu|, \mu = \frac{1}{n} \tag{2}$$

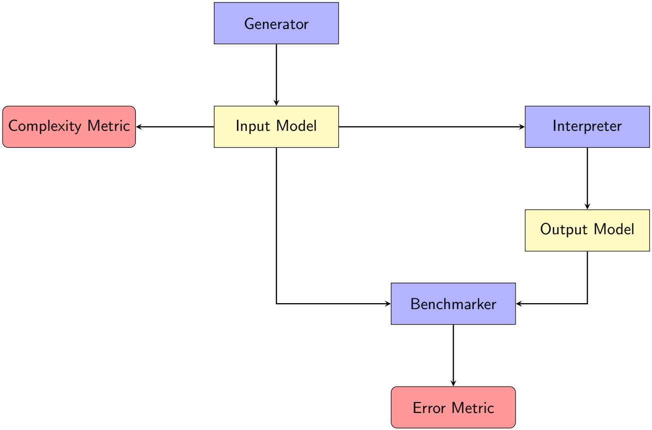
Error

Error

$$E(I, O) = \sum_{i=1}^{n} \sum_{j=1}^{n} |P_{I_{i,j}} - P_{O_{i,j}}|, \tag{3}$$

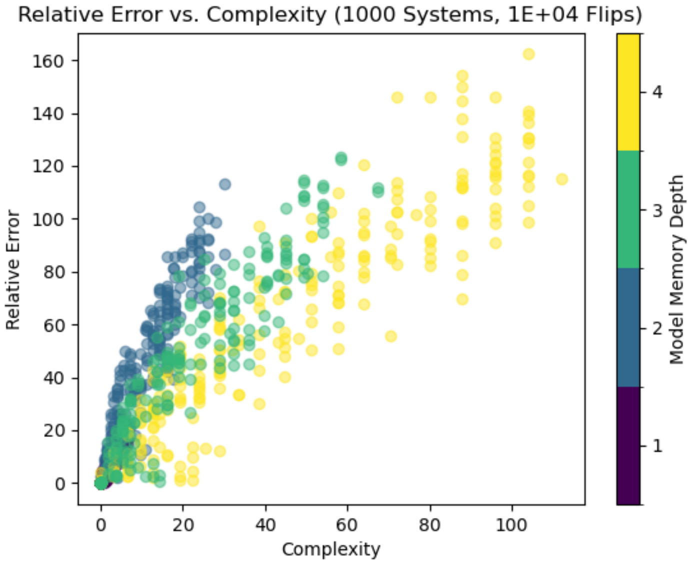
- The sum of the absolute differences for all probabilities across the distributions for all states on both systems.

Core Pipeline



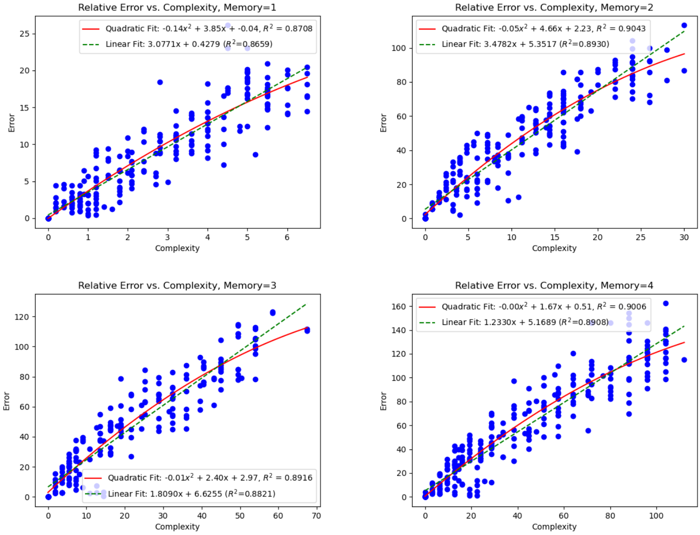
Interpreter: pre-informed of the transition rules and iterates over the data to calculate probability distributions for each state by building a histogram.

Results



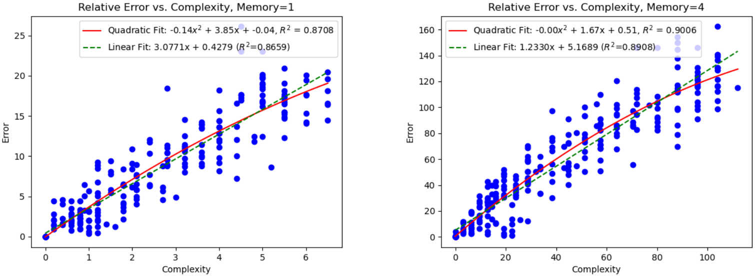
Relative Error vs. Complexity (1000 Systems, 1E+04 Flips)
Relative Error
0
20
40
60
80
100
120
140
160
Complexity
0
20
40
60
80
100
Model Memory Depth
1
2
3
4

Results



Relative Error vs. Complexity, Memory=1
Quadratic Fit: $-0.14x^2 + 3.85x + -0.04$, $R^2 = 0.8708$
Linear Fit: $3.0771x + 0.4279$ (R^2=0.8659)
Error
Complexity
Relative Error vs. Complexity, Memory=2
Quadratic Fit: $-0.05x^2 + 4.66x + 2.23$, $R^2 = 0.9043$
Linear Fit: $3.4782x + 5.3517$ (R^2=0.8930)
Error
Complexity
Relative Error vs. Complexity, Memory=3
Quadratic Fit: $-0.01x^2 + 2.40x + 2.97$, $R^2 = 0.8916$
Linear Fit: $1.8090x + 6.6255$ (R^2=0.8821)
Error
Complexity
Relative Error vs. Complexity, Memory=4
Quadratic Fit: $-0.00x^2 + 1.67x + 0.51$, $R^2 = 0.9006$
Linear Fit: $1.2330x + 5.1689$ (R^2=0.8908)
Error
Complexity

Results (Memory 1)



Slope at $m = 1$ is about 3, slope at $m = 4$ is about 1.2

Uncertainties

- Error measurements have inherent uncertainty, but should have limited influence due to the quantity of data
- Auto-generated data, and arbitrary quantities raise doubts over validity of results, these questions could be answered by researching with 'real' datasets
- Further tests over larger ranges of complexity are necessary to further validate our findings

Conclusion

- As expected, there is a positive/direct relationship between complexity and error.
- Model relationship cannot be determined due to our arbitrary quantities
- Memory influences the range of error for a given range of complexity

Questions!

Example Complexity and Variance

$$P = \begin{bmatrix} P_A \\ P_B \end{bmatrix} = \begin{bmatrix} 0.2 & 0.8 \\ 0.7 & 0.3 \end{bmatrix}$$

$T =$

Sequence	T_A	T_B
(0,0)	A	B
(0,1)	A	A
(1,0)	B	A
(1,1)	B	B

- $n = 2$ (width/height of P) and $m = 2$ (length of all sequences).
- As $n = 2$, $\mu = \frac{1}{2} = 0.5$.
  - $v = |0.2 - 0.5| + |0.8 - 0.5| + |0.7 - 0.5| + |0.3 - 0.5| = 1.0$
  - $c = nvm^2 = (2)(1)(2)^2 = 8$

Error Example

$$P_1 = \begin{bmatrix} 0.2 & 0.8 \\ 0.7 & 0.3 \end{bmatrix} \qquad P_2 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$$|P_1 - P_2| = \begin{bmatrix} |0.2 - 0.5| & |0.8 - 0.5| \\ |0.7 - 0.5| & |0.3 - 0.5| \end{bmatrix} = \begin{bmatrix} 0.3 & 0.3 \\ 0.2 & 0.2 \end{bmatrix}$$

$$\sum |P_1 - P_2| = 0.3 + 0.3 + 0.2 + 0.2 = \mathbf{1.0}$$

Interpreter

$n = 2, m = 2$

Dataset: 0,1,0,0,1,0,1,1,0,0

0 1	1 0	0 0	1 0	1 1	0 0

Histogram

Sequence	Count
(0,1)	2
(1,0)	2
(0,0)	2
(1,1)	1

Probability Distribution

Sequence	Probability
(0,1)	$2/7 \approx 0.286$
(1,0)	$2/7 \approx 0.286$
(0,0)	$2/7 \approx 0.286$
(1,1)	$1/7 \approx 0.143$