NKS2007 WolframScience CONFERENCE

An Alternative Approach to Anticipatory Reinforcement Learning

Alastair Hewitt

Harvard Extension School

Nondeterministic Information Systems

Model

Model consists of a set of *n* indicators – represents a finite state machine with 2^n states.

♦ Indicators

Each indicator is a Bernoulli trial measuring a specific *binary* condition: *"moving up"*, *"approaching point"*, *"within 5 units of edge".*

Model updated by performing measurements in sequence – state changes every time a single indicator changes – "*moving up to approach point*" is two state changes.

♦ Desirability

Each indicator transition is given a desirability *d* from 0→1, where zero is least desirable.

- ♦ Desirability represents intended frequency of event.
- ♦ *d*=0.5 is *equivalent* desirability, not an *ambiguous* choice.
- ♦ Opposite transition with desirability $d' = 1-d$

Strategies

Selection strategy uses prediction strategy to achieve goal of *equilibrium*.

♦ Prediction

Stores information about cause and effect correlations – a strong correlation indicates predictable behavior and the availability of a reliable prediction strategy.

♦ Selection

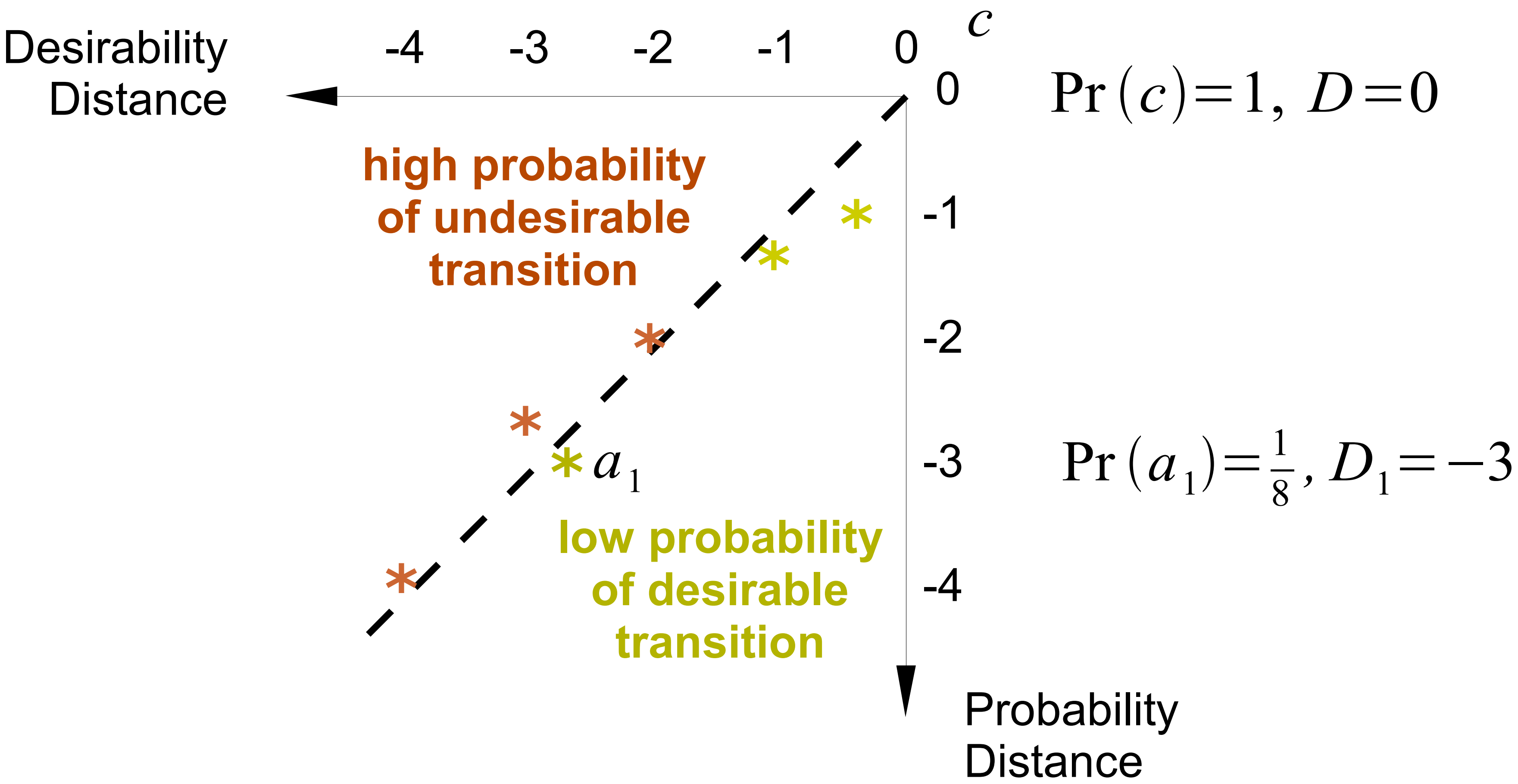
Searches prediction strategy to determine the best choice of action – constrained by the level of *doubt* involved in the predictions being used.

♦ Goal

A *tally* is made for transition against the action currently selected – the probability of this correlation is simply the tally divided by the total.
The goal is to reach an equilibrium, where the ratio of probability and desirability equals 1.

Equilibrium

Choices are made to minimize the absolute difference between probability and desirability *distance D*, where *D* is the base 2 logarithm of probability (or desirability) [SZIJÁRTÓ GRÖGER KALLÓS 2002].



high probability of undesirable transition

low probability of desirable transition

$$\Pr(c)=1,\ D=0$$

$$\Pr(a_1)=\tfrac{1}{8},\ D_1=-3$$

Predictability and Doubt

Doubt is a measure of uncertainty and defined as the logarithm of the information *redundancy* [Shannon 1948]. The *predictability* of the system is the inverse of doubt – ranging from 0 *(random)* to infinity *(deterministic)*.

- ♦ **Redundancy**

$$R = 1 - H_{rel} \qquad H_{rel} = \frac{H}{H_{max}}$$

- ♦ **Entropy**

$$H = -\sum_{i=1}^{n} p_i \log p_i \qquad H_{max} = -\log n$$

- ♦ **Doubt**

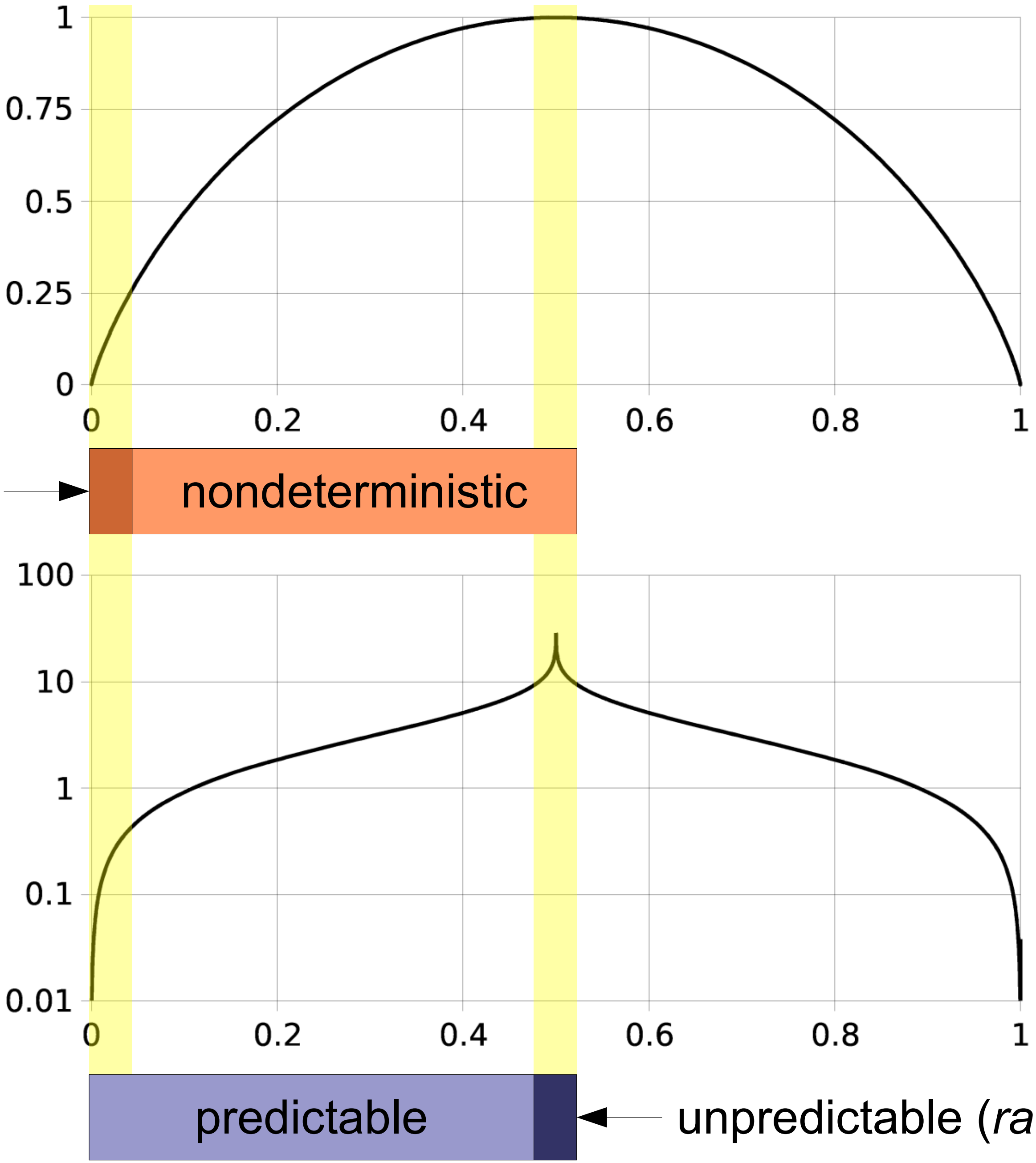
$$U = \log \log n - \log\left(\log n + \sum_{i=1}^{n} p_i \log p_i\right)$$

Entropy

$$\{0 < H_{rel} \leq 1\}$$

Doubt

$$\{0 < U < \infty\}$$



deterministic

nondeterministic

predictable

unpredictable (*random*)

Elementary System – 2D CA

- **2 choices** : turn (**T**)
no action (**N**)

- **1 indicator** *(2 states)* : *"going in right direction"* (**1**)
"going in wrong direction" (**0**)

$$\left.\begin{array}{l} d_1=0.75\ ,\ D_1=-0.415 \\ d_0=1-d_1=0.25\ ,\ D_0=-2 \end{array}\right\}\ U_E=2.4$$

Expected doubt at equilibrium

- **State Table**

state table **1**

Actions

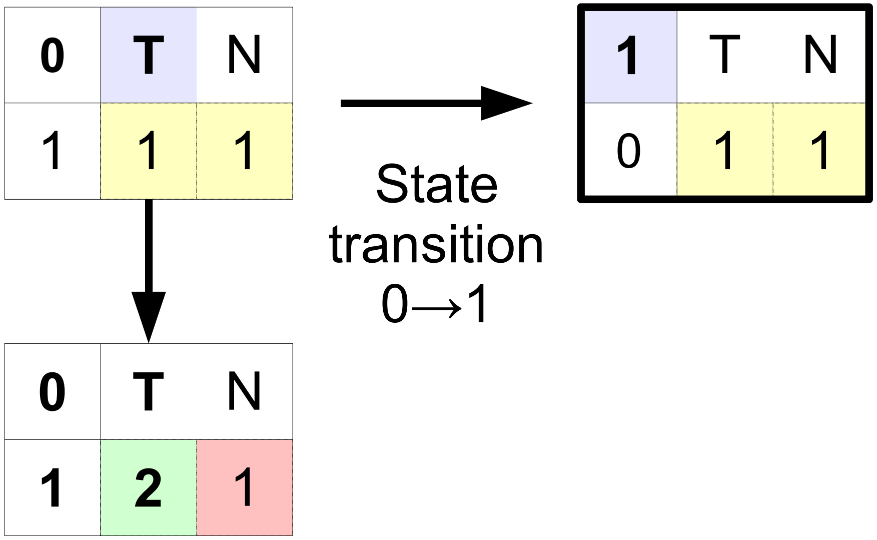
Transition 1→0

1	T	N
0	2	1

Best choice

Elementary System step 1

$$\left.\begin{array}{l} p_T=\frac{1}{2}\ ,\ D_T=-1\ ,\ |E_T|=0.585 \\ p_N=\frac{1}{2}\ ,\ D_N=-1\ ,\ |E_N|=0.585 \end{array}\right\}\ U_0=\infty$$



Can't decide.
Select default:
Action **T**

0	**T**	N
1	1	1

State
transition
0→1

1	T	N
0	1	1

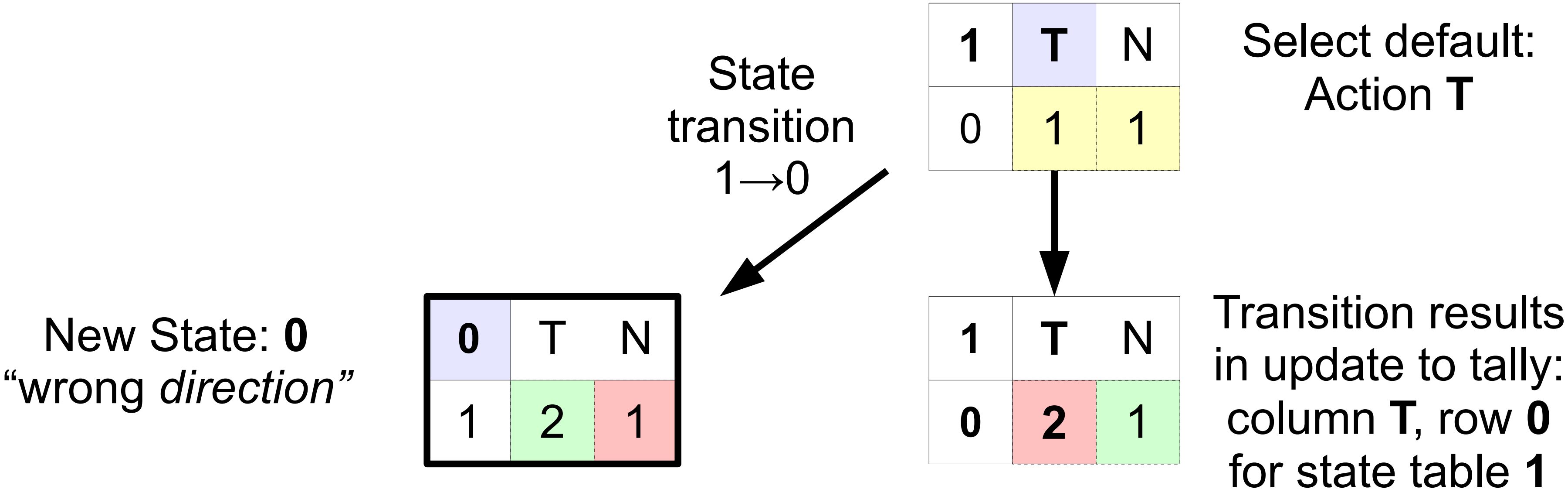
New State: **1**
"right direction"

Transition results
in update to tally:
column **T**, row **1**
for state table **0**

0	**T**	N
1	**2**	1

Elementary System step 2

$$\left.\begin{array}{l} p_T=\frac{1}{2}\ ,\ D_T=-1\ ,\ |E_T|=1 \\ p_N=\frac{1}{2}\ ,\ D_N=-1\ ,\ |E_N|=1 \end{array}\right\}\ U_1=\infty$$



1	**T**	N
0	1	1

Select default:
Action **T**

State
transition
1→0

1	**T**	N
0	**2**	1

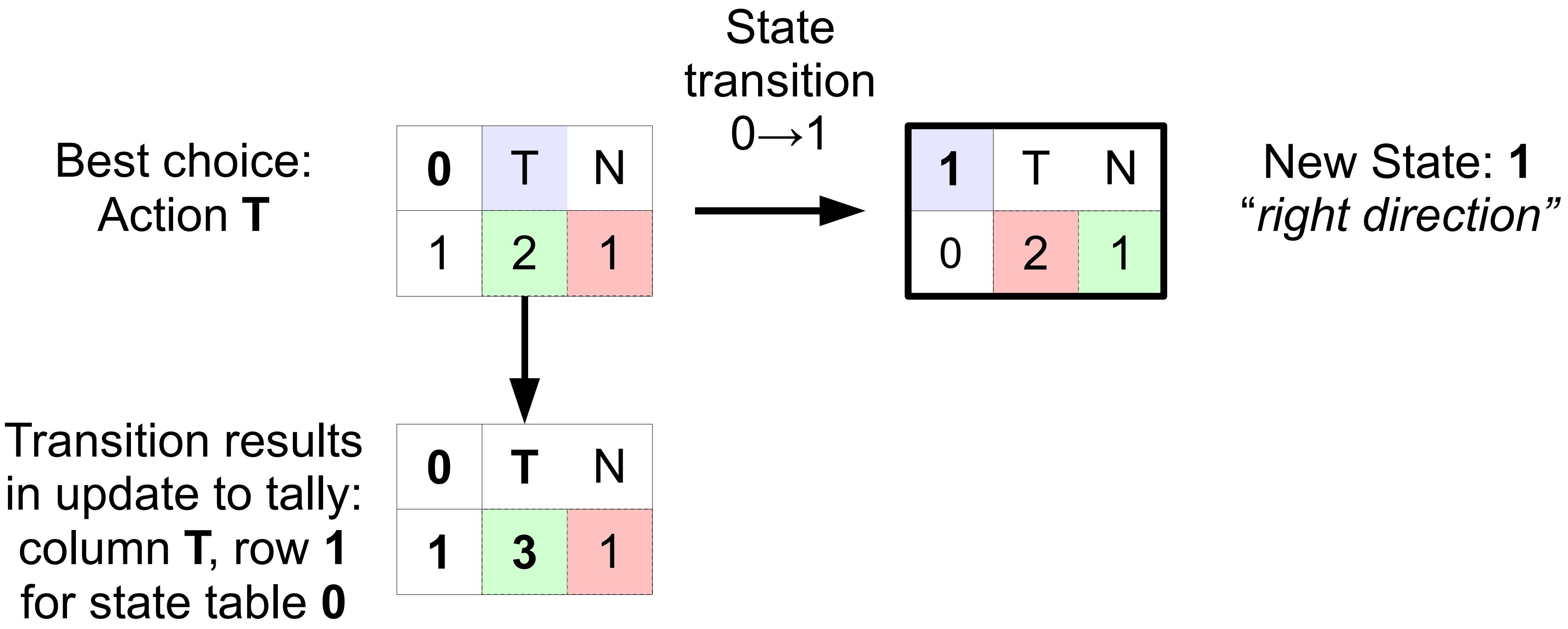
Transition results
in update to tally:
column **T**, row **0**
for state table **1**

New State: **0**
“wrong *direction*”

0	T	N
1	2	1

Elementary System step 3

$$\left.\begin{array}{l} p_T=\frac{2}{3}\ ,\ D_T=-0.585\ ,\ |E_T|=0.17 \\ p_N=\frac{1}{3}\ ,\ D_N=-1.585\ ,\ |E_N|=1.17 \end{array}\right\} U_0=3.6$$



Best choice:
Action **T**

0	T	N
1	2	1

State
transition
0→1

1	T	N
0	2	1

New State: **1**
“right direction”

Transition results
in update to tally:
column **T**, row **1**
for state table **0**

0	**T**	N
1	**3**	1

Elementary System step 4

$$\left.\begin{array}{l} p_T=\frac{2}{3}\ ,\ D_T=-0.585\ ,\ |E_T|=1.415 \\ p_N=\frac{1}{3}\ ,\ D_N=-1.585\ ,\ |E_N|=0.415 \end{array}\right\} U_1=3.6$$

Selecting: **no action**
results in no further
state transitions

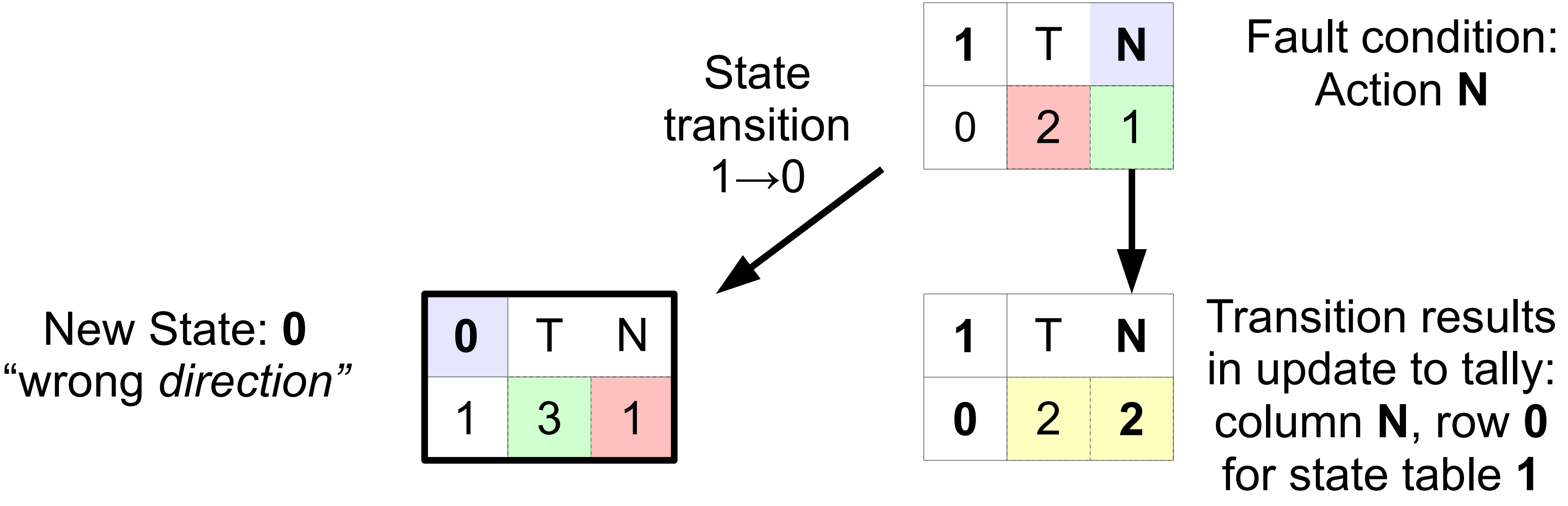
1	T	**N**
0	2	1

Best choice:
Action **N**
“right direction”

0	T	N
1	3	1

Elementary System fault

Introduce an *obstacle* that deflects agent into *wrong direction* whilst *no action* was selected. Results in a fault condition, where learning is erased and doubt is returned to maximum.



1	T	**N**
0	2	1

Fault condition:
Action **N**

State
transition
1→0

0	T	N
1	3	1

New State: **0**
"wrong *direction*"

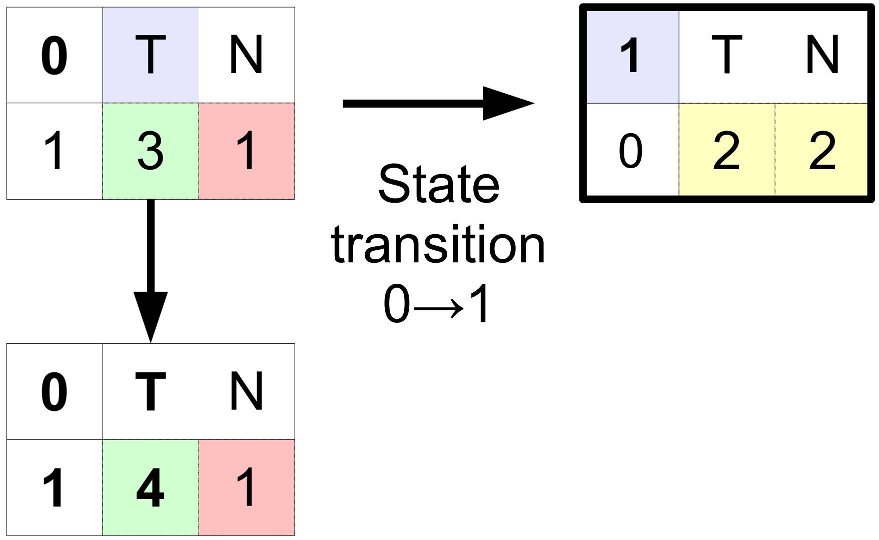
1	T	**N**
0	2	**2**

Transition results
in update to tally:
column **N**, row **0**
for state table **1**

Elementary System step 5

$$\left.\begin{array}{l} p_T=\frac{3}{4}\;,\; D_T=-0.415\;,\; |E_T|=0 \\ p_N=\frac{1}{4}\;,\; D_N=-2\;,\; |E_N|=1.585 \end{array}\right\} U_0=2.4$$

Best choice:
Action **T**



0	T	N
1	3	1

State transition 0→1

1	T	N
0	2	2

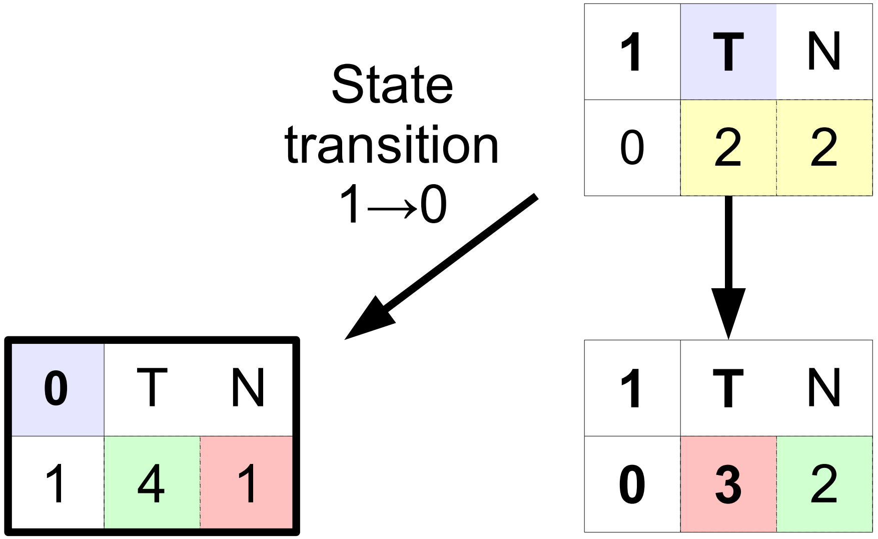
New State: **1**
“right direction”

Transition results in update to tally: column **T**, row **1** for state table **0**

0	**T**	N
1	**4**	1

Elementary System step 6

$$\left.\begin{array}{l} p_T=\frac{1}{2}\ ,\ D_T=-1\ ,\ |E_T|=1 \\ p_N=\frac{1}{2}\ ,\ D_N=-1\ ,\ |E_N|=1 \end{array}\right\}\ U_1=\infty$$



State transition 1→0

1	**T**	N
0	2	2

Select default:
Action **T**

1	**T**	N
0	**3**	2

Transition results in update to tally: column **T**, row **0** for state table **1**

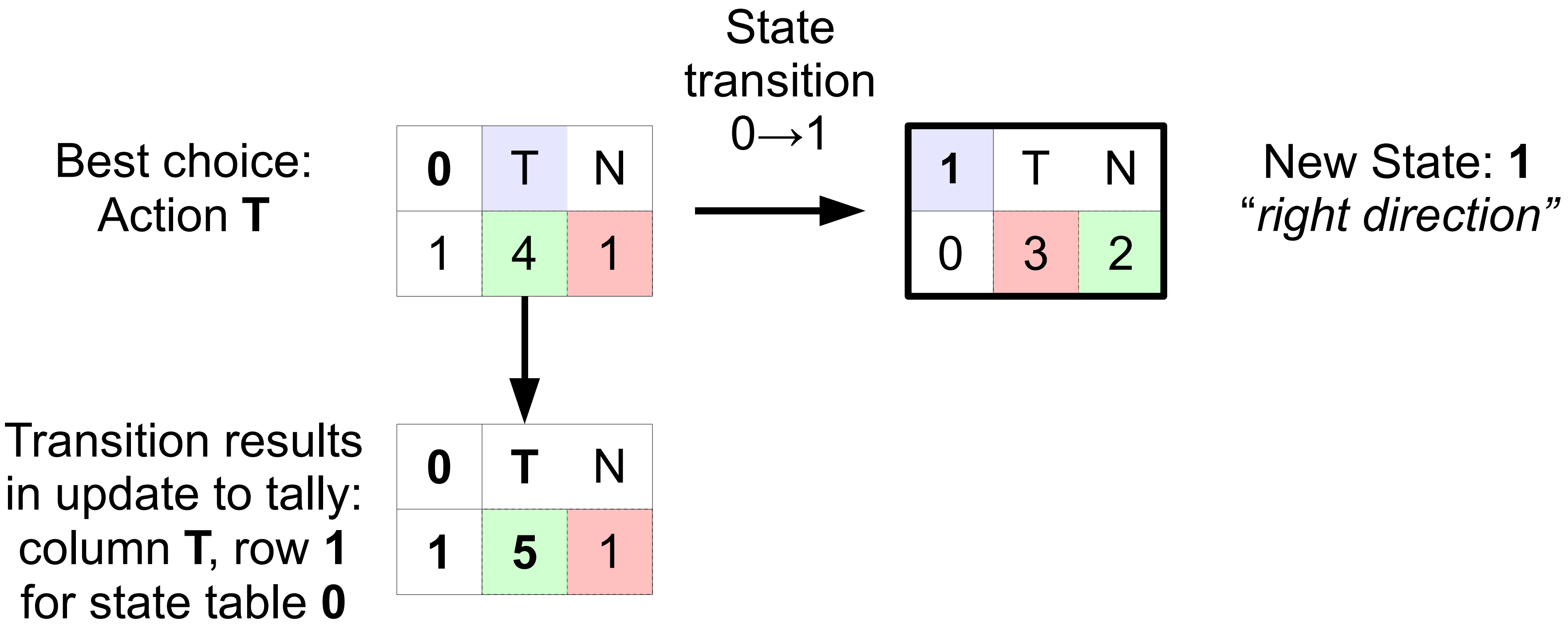
New State: **0**
“wrong *direction*”

0	T	N
1	4	1

Elementary System

step 7

$$\left.\begin{array}{l} p_T=\frac{4}{5}\ ,\ D_T=-0.322\ \ ,\ |E_T|=0.093 \\ p_N=\frac{1}{5}\ ,\ D_N=-2.322\ ,\ |E_N|=1.907 \end{array}\right\}\ U_0=1.5$$



Best choice:
Action **T**

0	T	N
1	4	1

State
transition
0→1

1	T	N
0	3	2

New State: **1**
"*right direction*"

Transition results
in update to tally:
column **T**, row **1**
for state table **0**

0	**T**	N
1	**5**	1

Elementary System step 8

$$\left.\begin{array}{l} p_T=\frac{3}{5}\ ,\ D_T=-0.737\ ,\ |E_T|=1.263 \\ p_N=\frac{2}{5}\ ,\ D_N=-1.322\ ,\ |E_N|=0.678 \end{array}\right\}\ U_1=5.1$$

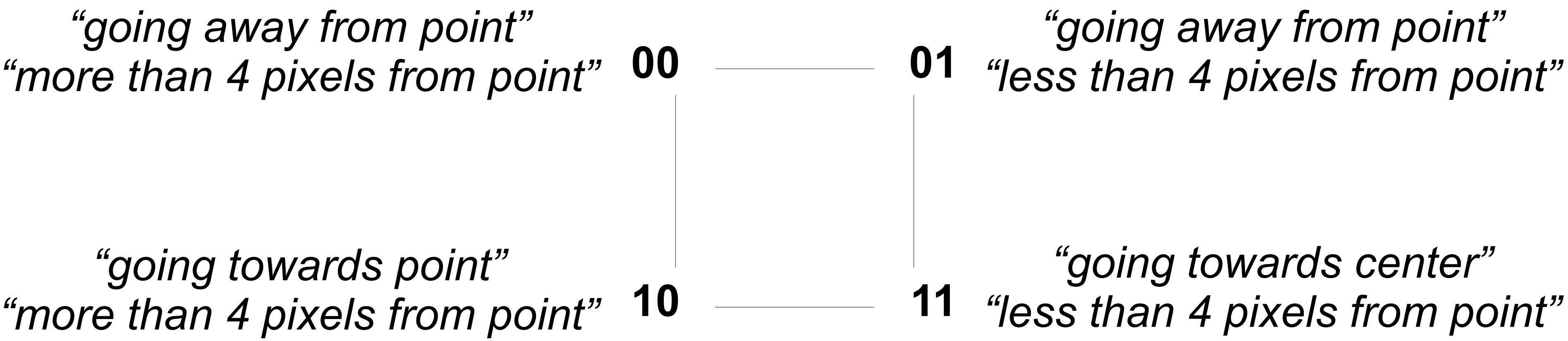
Selecting: **no action**
results in no further
state transitions

1	T	N
0	**3**	2

Best choice:
Action **N**
"right direction"

0	T	N
1	**5**	1

Two Indicator System (*demo*)



"going away from point"
"more than 4 pixels from point"
00

01
"going away from point"
"less than 4 pixels from point"

"going towards point"
"more than 4 pixels from point"
10

11
"going towards center"
"less than 4 pixels from point"

♦ State Table

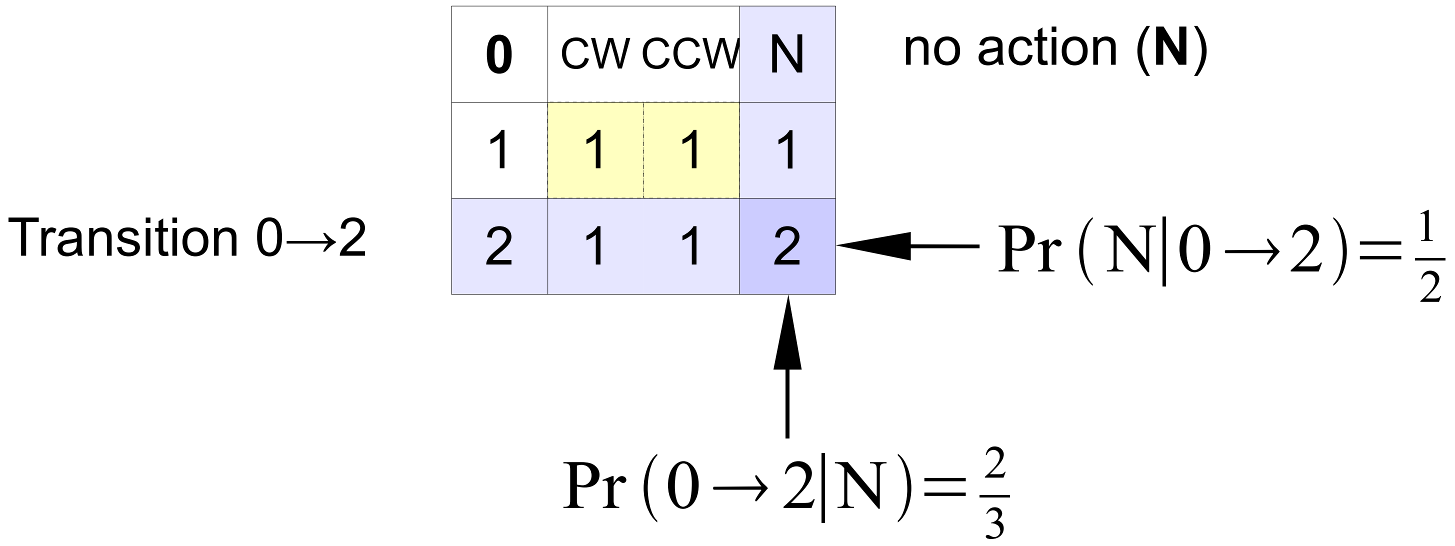
state table **0**

	0	CW	CCW	N
Transition 0→1	1	1	1	1
Transition 0→2	2	1	1	2

turn clockwise (**CW**)
turn counter clockwise (**CCW**)
no action (**N**)

Best choice

Design Features – Probability



no action (**N**)

Transition 0→2

0	CW	CCW	N
1	1	1	1
2	1	1	2

$$\Pr\left(N \middle| 0 \rightarrow 2\right) = \frac{1}{2}$$

$$\Pr\left(0 \rightarrow 2 \middle| N\right) = \frac{2}{3}$$

♦ Column Sum

“what is probability of transition given action” - conventional cause and effect correlation, but equals 1 for single transition table (one row).

♦ Row Sum

“what is probability of choosing action given transition” - cause follows effect, agent forces system to reflect own desires.

Other Design Features

♦ Choice Criteria

To determine best choice of action for a single table:

- ♦ Equilibrium distance calculated from probability via row sum.
- ♦ Total distance calculated using RMS sum for column.
- ♦ Smallest sum represents best choice of action.

♦ Search Algorithm

Search multiple adjacent tables to anticipate:

- ♦ Largest tally in column assumed to be next transition given action.
- ♦ Find shortest deviation from equilibrium following paths.
- ♦ Depth first search constrained by *accumulated* doubt.

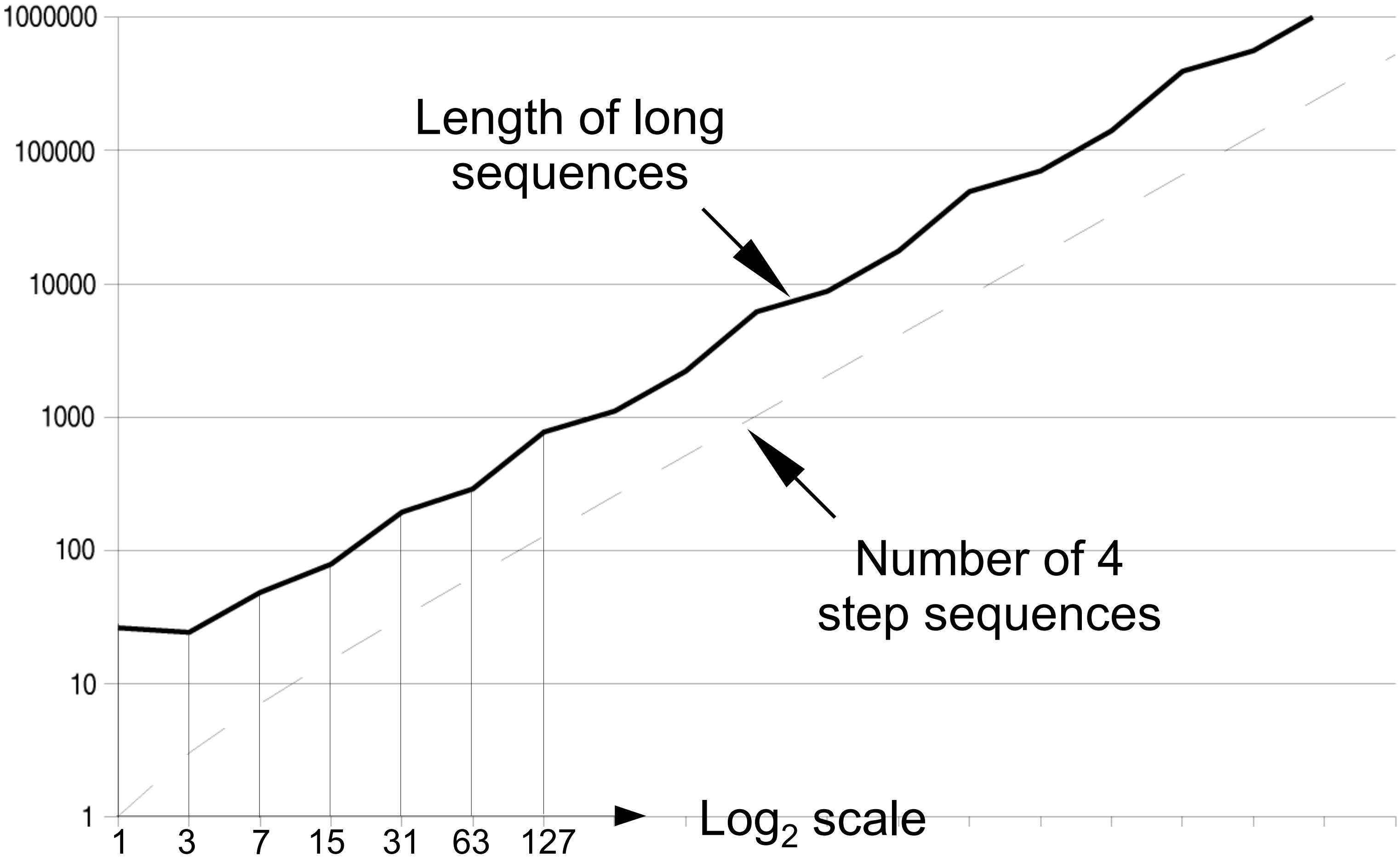
♦ Integer Math

High precision not necessary, but accuracy does effect behavior.

- ♦ Rounding to integers reveals repetition faster.
- ♦ Logarithms calculated as integers by counting leading zeros.
- ♦ Probabilities are treated as rational numbers.

One Agent Results

Collisions with central point evolve by 2^n sequences of 4 steps, then one sequence of $2^n+f(n)$ steps before next collision – develops repetitive behavior after 2444 steps.



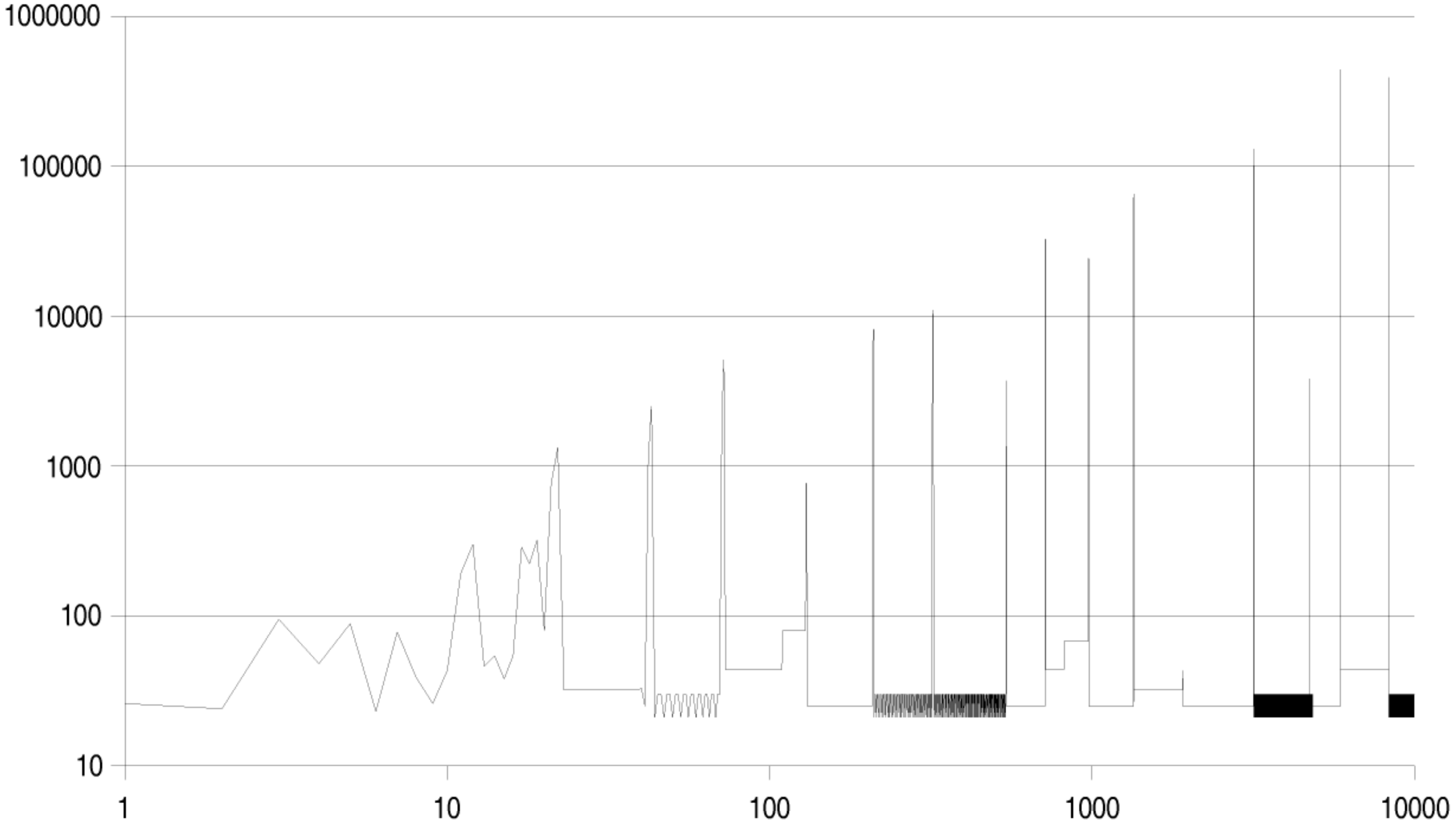
Length of long sequences

Number of 4 step sequences

Log_2 scale

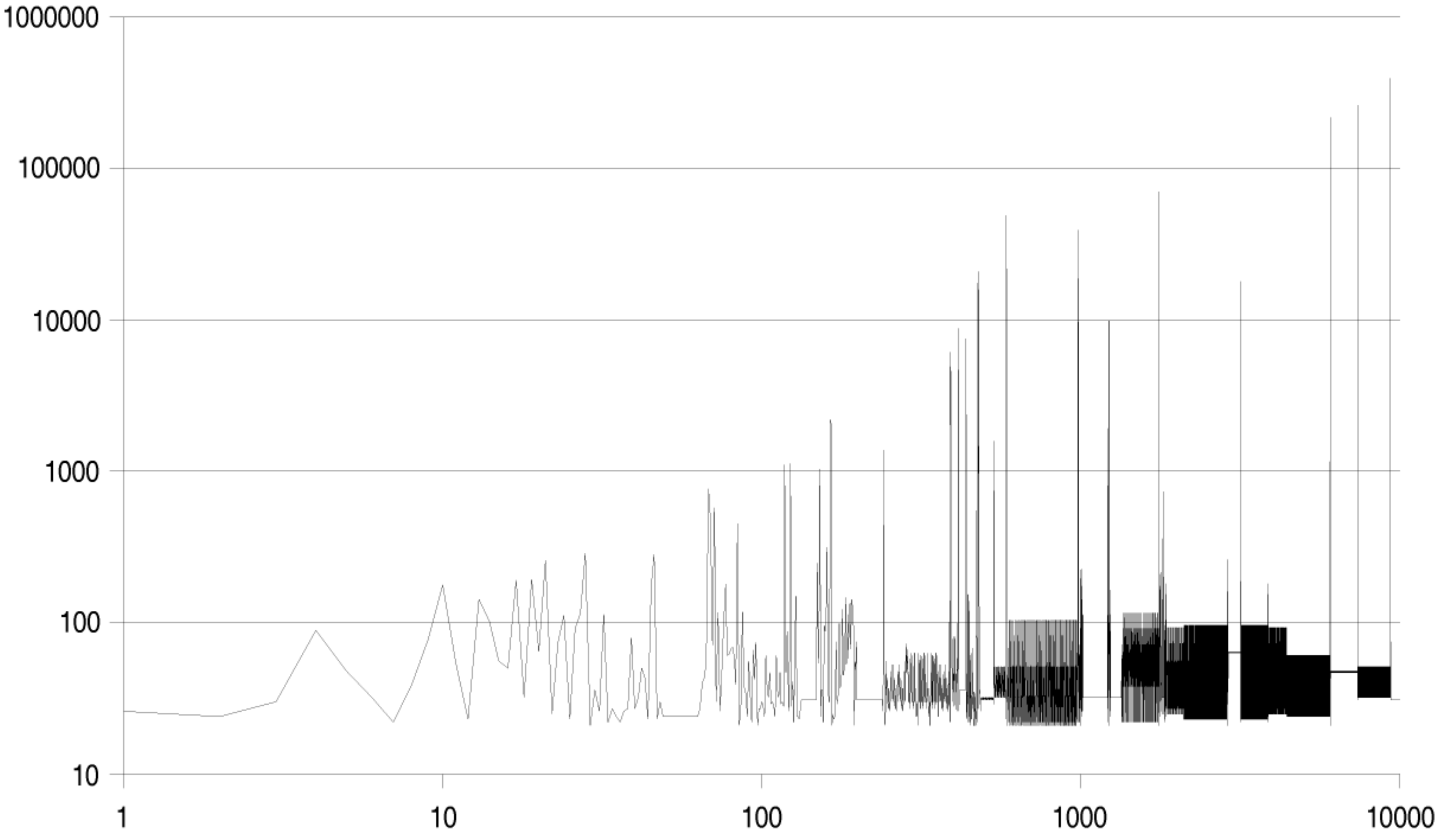
Two Agent Results

Steps between collisions evolve less predictably, but still structured. Number of short sequences and length of long sequences increase as system evolves.



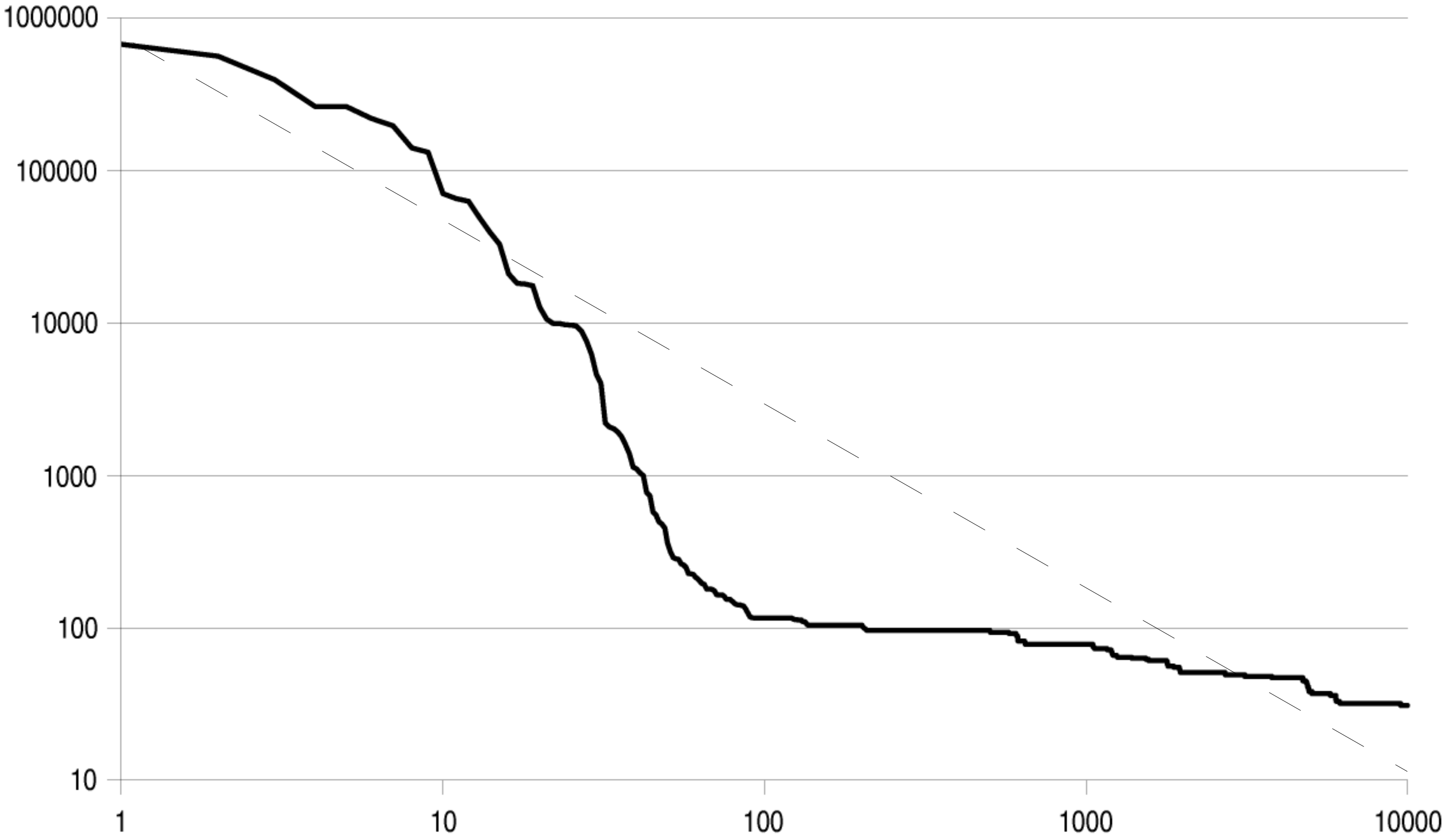
Three Agent Results

Complexity increases as number of agents increases, length of repetitive behavior still increases as system evolves.



Three Agent Observation

Distribution of sequence length exhibits some features of a power law – *weak* indicator of randomness (pink noise, Brownian motion).



References

Sutton, R., Barto, G.: Reinforcement Learning, An Introduction, MIT Press, (1998).

Anticipative Reinforcement Learning, Maire, F., Proceedings of the 9th International Conference on Neural Information Processing, Vol. 3, PP. 1428- 1432 (2002)

Shannon, C. E., A Mathematical Theory of Communication, The Bell System Technical Journal, XXVII (1948).

Szijártó M., Gröger, D., Kallós G.: A Distance Model for Safety-Critical Systems, Periodica Polytechnica Ser. El. Eng. Vol. 45, No. 2, PP. 109-118 (2001).