

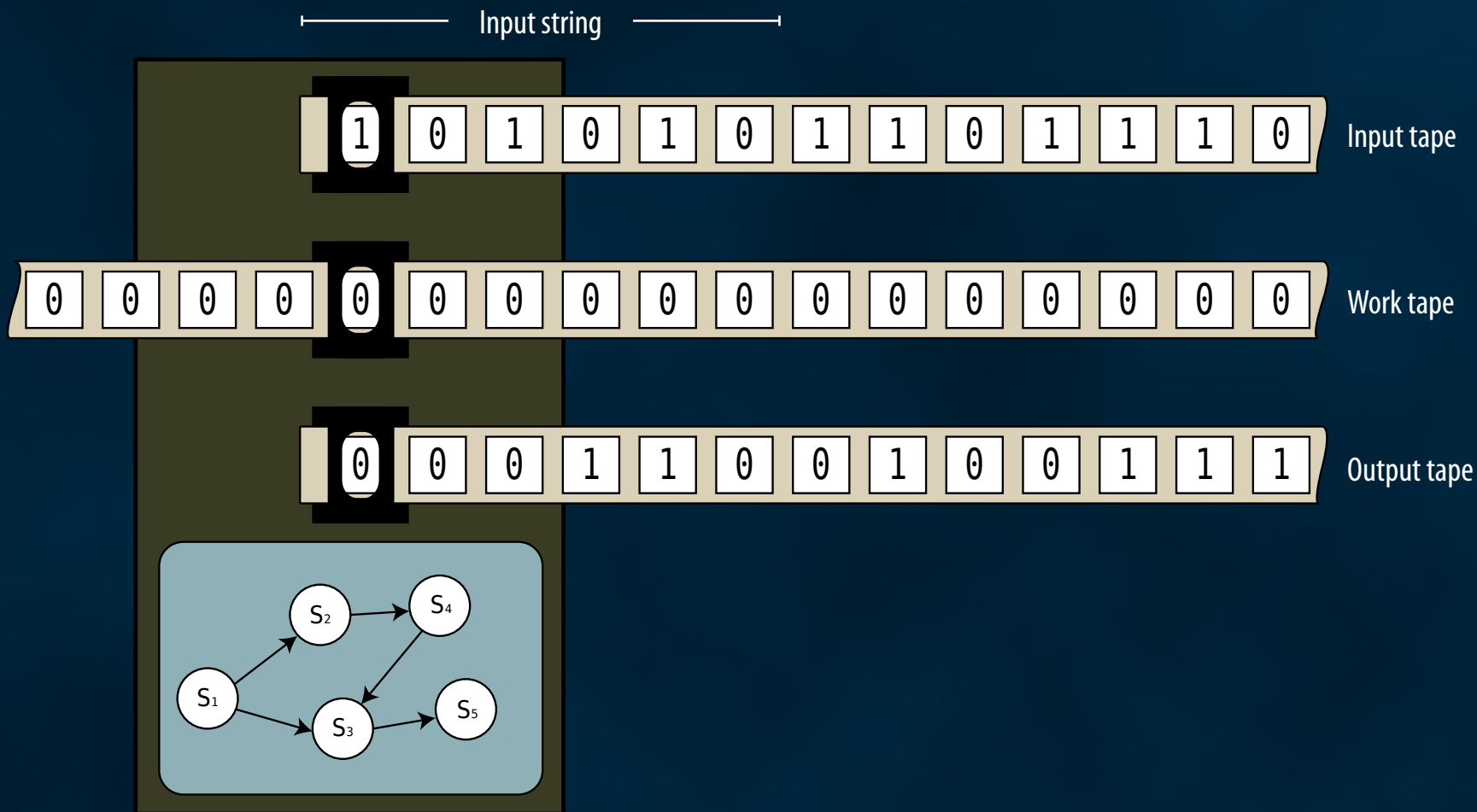
Minimum Message Length and Kolmogorov Complexity

C. S. Wallace and D. L. Dowe



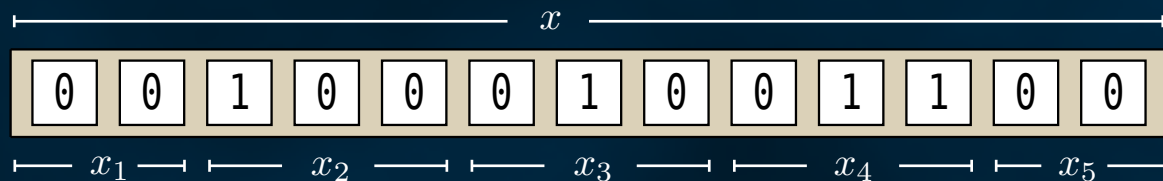
Overview

Turing Machines



Data & Hypotheses

Data string x is a representation of observational data from a real world phenomenon



$$L = \{00, 100, 010, 011\}$$

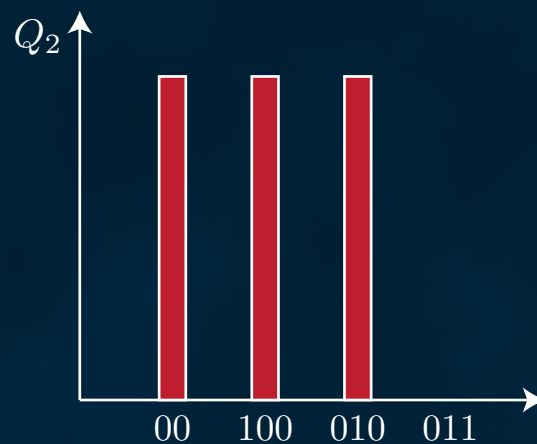
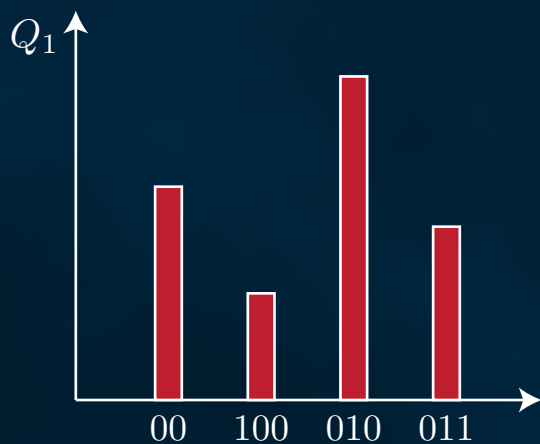
- “Sentences” $x_i \in L$, where L is a prefix-free set (data “language”)
- Distinct sentences represent distinct real-world facts
- Sentences are conditionally independent given full knowledge of the phenomenon
- Strings are invariant to sentence permutation

Data & Hypotheses

Hypothesis Q is a (computable) probability distribution over L

Conditional independence of sentences implies

$$x = x_1 \dots x_n \Rightarrow Q(x) = Q(x_1) \times \dots \times Q(x_n)$$





Two-part encoding

How do we acquire a hypothesis-based encoding of data in the Algorithmic Complexity framework?

Two-part encoding

How do we acquire a hypothesis-based encoding of data in the Algorithmic Complexity framework?

Idea: Use conditional Kolmogorov complexity

$$K_T(x \mid y) = \min\{l(p) \mid T(\langle y, p \rangle) = x\}$$

and interpret y as hypothesis and x as data.

Two-part encoding

How do we acquire a hypothesis-based encoding of data in the Algorithmic Complexity framework?

Idea: Use conditional Kolmogorov complexity

$$K_T(x \mid y) = \min\{l(p) \mid T(\langle y, p \rangle) = x\}$$

and interpret y as hypothesis and x as data.

Corresponding conditional algorithmic probability

$$P_T(x \mid y) = 2^{-K_T(x \mid y)}$$

Two-part encoding

How do we acquire a hypothesis-based encoding of data in the Algorithmic Complexity framework?

Idea: Use conditional Kolmogorov complexity

$$K_T(x \mid y) = \min\{l(p) \mid T(\langle y, p \rangle) = x\}$$

and interpret y as hypothesis and x as data.

Corresponding conditional algorithmic probability

$$P_T(x \mid y) = 2^{-K_T(x \mid y)}$$

Problem: Probability can never be 0, i.e. Popper-falsification not possible, because

$$K(x \mid y) < K(x) + O(1) \Rightarrow P_K(x \mid y) > P_K(x) + O(1)$$

Why?

Two-part encoding

How do we acquire a hypothesis-based encoding of data in the Algorithmic Complexity framework?

Idea: Use conditional Kolmogorov complexity

$$K_T(x \mid y) = \min\{l(p) \mid T(\langle y, p \rangle) = x\}$$

and interpret y as hypothesis and x as data.

Corresponding conditional algorithmic probability

$$P_T(x \mid y) = 2^{-K_T(x \mid y)}$$

Problem: Probability can never be 0, i.e. Popper-falsification not possible, because

$$K(x \mid y) < K(x) + O(1) \Rightarrow P_K(x \mid y) > P_K(x) + O(1)$$

Why? Hypothesis y acts as “extra info”, instead of assertively.

Two-part encoding

How do we acquire a hypothesis-based encoding of data in the Algorithmic Complexity framework?

Idea: Use conditional Kolmogorov complexity

$$K_T(x \mid y) = \min\{l(p) \mid T(\langle y, p \rangle) = x\}$$

and interpret y as hypothesis and x as data.

Corresponding conditional algorithmic probability

$$P_T(x \mid y) = 2^{-K_T(x \mid y)}$$

Problem: Probability can never be 0, i.e. Popper-falsification not possible, because

$$K(x \mid y) < K(x) + O(1) \Rightarrow P_K(x \mid y) > P_K(x) + O(1)$$

Why? Hypothesis y acts as “extra info”, instead of assertively.

Proposal: Have hypothesis be a prefix of input string p .
Force intended two-part encoding by imposing conditions on p .

Two-part encoding

Input p is an acceptable MML message encoding data string x , if

Two-part encoding

Input p is an acceptable MML message encoding data string x , if

1) $T(p) = x$

p encodes x

Two-part encoding

Input p is an acceptable MML message encoding data string x , if

1) $T(p) = x$

p encodes x

2) $l(p) < l(x)$

some compression is achieved

Two-part encoding

Input p is an acceptable MML message encoding data string x , if

1) $T(p) = x$

p encodes x

2) $l(p) < l(x)$

some compression is achieved

3) $p = qr$

two-part encoding

Two-part encoding

Input p is an acceptable MML message encoding data string x , if

1) $T(p) = x$

p encodes x

2) $l(p) < l(x)$

some compression is achieved

3) $p = qr$

two-part encoding

4) $T(q) = \epsilon$

hypothesis q is does not determine data

Two-part encoding

Input p is an acceptable MML message encoding data string x , if

1) $T(p) = x$

p encodes x

2) $l(p) < l(x)$

some compression is achieved

3) $p = qr$

two-part encoding

4) $T(q) = \epsilon$

hypothesis q is does not determine data

5) $T_q(rs) = xT_q(s)$

reading r does not alter the state of T

Two-part encoding

Input p is an acceptable MML message encoding data string x , if

1) $T(p) = x$

p encodes x

2) $l(p) < l(x)$

some compression is achieved

3) $p = qr$

two-part encoding

4) $T(q) = \epsilon$

hypothesis q is does not determine data

5) $T_q(rs) = xT_q(s)$

reading r does not alter the state of T

6) $l(r) < K_T(x)$

hypothesis q is "significant"

Two-part encoding

Input p is an acceptable MML message encoding data string x , if

$$1) \quad T(p) = x$$

p encodes x

$$2) \quad l(p) < l(x)$$

some compression is achieved

$$3) \quad p = qr$$

two-part encoding

$$4) \quad T(q) = \epsilon$$

hypothesis q is does not determine data

$$5) \quad T_q(rs) = xT_q(s)$$

reading r does not alter the state of T

$$6) \quad l(r) < K_T(x)$$

hypothesis q is "significant"

$$7) \quad x = x_1 \dots x_n \Rightarrow \begin{cases} r = r_1 \dots r_n \\ T_q(r_i) = x_i, \quad i = 1 \dots n \end{cases}$$

conditionally independent sentences

Two-part encoding

Input p is an acceptable MML message encoding data string x , if

$$1) \quad T(p) = x$$

p encodes x

$$2) \quad l(p) < l(x)$$

some compression is achieved

$$3) \quad p = qr$$

two-part encoding

$$4) \quad T(q) = \epsilon$$

hypothesis q is does not determine data

$$5) \quad T_q(rs) = xT_q(s)$$

reading r does not alter the state of T

$$6) \quad l(r) < K_T(x)$$

hypothesis q is "significant"

$$7) \quad x = x_1 \dots x_n \Rightarrow \begin{cases} r = r_1 \dots r_n \\ T_q(r_i) = x_i, \quad i = 1 \dots n \end{cases}$$

conditionally independent sentences

$$8) \quad \begin{matrix} x' = x^{(1)}x^{(2)} \\ j' = j^{(1)}j^{(2)} \end{matrix} \Rightarrow \begin{matrix} T_q(j^{(1)}) = x^{(1)}, \quad j^{(1)} < K_T(x^{(1)}) \\ T_q(j^{(2)}) = x^{(2)}, \quad j^{(2)} < K_T(x^{(2)}) \end{matrix}$$

hypothesis q is "general"

Two-part encoding

Input p is an acceptable MML message encoding data string x , if

$$1) \quad T(p) = x$$

p encodes x

$$2) \quad l(p) < l(x)$$

some compression is achieved

$$3) \quad p = qr$$

two-part encoding

$$4) \quad T(q) = \epsilon$$

hypothesis q is does not determine data

$$5) \quad T_q(rs) = xT_q(s)$$

reading r does not alter the state of T

$$6) \quad l(r) < K_T(x)$$

hypothesis q is "significant"

$$7) \quad x = x_1 \dots x_n \Rightarrow \begin{cases} r = r_1 \dots r_n \\ T_q(r_i) = x_i, \quad i = 1 \dots n \end{cases}$$

conditionally independent sentences

$$8) \quad \begin{matrix} x' = x^{(1)}x^{(2)} \\ j' = j^{(1)}j^{(2)} \end{matrix} \Rightarrow \begin{matrix} T_q(j^{(1)}) = x^{(1)}, \quad j^{(1)} < K_T(x^{(1)}) \\ T_q(j^{(2)}) = x^{(2)}, \quad j^{(2)} < K_T(x^{(2)}) \end{matrix}$$

hypothesis q is "general"

$$9) \quad \text{No prefix of } q \text{ satisfies all the above conditions}$$

all of q is required