CS 483 - Introduction to the Theory of Computation

Lecture 18

Kolmogorov Complexity

Consider the following two strings:

$$x = 01010101010101010101$$

$$y = 11001101000010101100$$

y is clearly more "complex" than x.

How do we formally capture this intuition of complexity?

Kolmogorov Complexity

For $x \in \{0,1\}^*$, the Kolmogorov complexity (or descriptive complexity) $K(x)$ is defined as

$$\min \{ |(M, w)| \mid M(w) = x \}$$

where M is a Turing machine, and $|(M, w)|$ is the length of the binary encoding of (M, w).

The encoding of (M, w) achieving the length $K(x)$ is called the minimum description of x.

Basic Facts about Kolmogorov Complexity

Theorem. For every computable function $f: \{0,1\}^* \to \{0,1\}^*$, let

$$K_f(x) = \min \{ |y| \mid f(y) = x \}.$$

Then there exists a constant c_f such that for every x,

$$K(x) \leq K_f(x) + c_f.$$

In other words, Kolmogorov complexity indicates the shortest length of descriptions no matter how you translate them.

Basic Facts about Kolmogorov Complexity

Theorem. For every computable function $f: \{0,1\}^* \to \{0,1\}^*$, let

$$K_f(x) = \min \{ |y| \mid f(y) = x \}.$$

Then there exists a constant c_f such that for every x,

$$K(x) \leq K_f(x) + c_f.$$

Proof. Suppose M computes f, and $K_f(x) = |y|$ where $M(y) = f(y) = x$. Then $K(x) \leq |(M, y)| = |y| + c_f$, where c_f is the length of the encodings for $(M,)$.

Basic Facts about Kolmogorov Complexity

Theorem. For every computable function $f:\{0,1\}^* \to \{0,1\}^*$, let

$$K_f(x) = \min \{ |y| \mid f(y) = x \}.$$

Then there exists a constant c_f such that for every x,

$$K(x) \leq K_f(x) + c_f.$$

Corollary. There is a constant c such that for every $x \in \{0,1\}^*$,

$$K(x) \leq |x| + c.$$

Basic Facts about Kolmogorov Complexity

Theorem. For every $n, c \in \mathbb{N}$, and a uniformly random $x \in \{0,1\}^n$,

$$\Pr[K(x) \leq |x| - c] < 2^{-c+1}.$$

Basic Facts about Kolmogorov Complexity

Theorem. For every $n, c \in \mathbb{N}$, and a uniformly random $x \in \{0,1\}^n$,

$$\Pr[K(x) \leq |x| - c] < 2^{-c+1}.$$

Proof. The number of strings with minimum description length at most $n - c$ is at most $2^0 + 2^1 + \cdots + 2^{n-c} < 2^{n-c+1}$, so

$$\Pr[K(x) \leq |x| - c] < \frac{2^{n-c+1}}{2^n}.$$

Basic Facts about Kolmogorov Complexity

Theorem. For every $n, c \in \mathbb{N}$, and a uniformly random $x \in \{0,1\}^n$,

$$\Pr[K(x) \leq |x| - c] < 2^{-c+1}.$$

Corollary. For every $n \in \mathbb{N}$, there exists $x \in \{0,1\}^n$ such that

$$K(x) \geq |x|.$$

Such x is called incompressible (as it cannot be described with shorter strings).

Note that the statement also holds for K_f.

Incompressibility

Can we decide if a string is incompressible?

Incompressibility

Can we decide if a string is incompressible?

If so, consider x_{100}, the first incompressible string of length 100.

But "the first incompressible string of length 100" is a short description for x_{100}!

Berry paradox: "The smallest natural number not definable in ten words."

Incompressibility

Theorem. $\textsc{Incomp} = \{ x \in \{0,1\}^* \mid K(x) \geq |x| \}$ is undecidable.

Proof. Assume M decides $\textsc{Incomp}$, construct $M'(n)$:

- For every $x \in \{0,1\}^n$:
  - Run $M(x)$. If $M(x) = 1$ then output x and halt.

Let x_n be the first incompressible string in $\{0,1\}^n$. Then there exists a constant c that

$$K(x_n) \leq |(M', n)| \leq \log_2 n + c,$$

Which contradicts $K(x_n) \geq n$ when n is large enough.

Applications of Incompressibility

Example 1: There are infinitely many prime numbers.

Applications of Incompressibility

Example 1: There are infinitely many prime numbers.

Assume there are finitely many prime numbers $p_1, \dots, p_k$. For every $x \in \{0,1\}^*$, assume $(1x)_2 = N = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$, then x can be described as

$$y = a_1 \# a_2 \# \cdots \# a_k$$

Applications of Incompressibility

Example 1: There are infinitely many prime numbers.

Assume there are finitely many prime numbers $p_1, \dots, p_k$. For every $x \in \{0,1\}^*$, assume $(1x)_2 = N = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$, then x can be described as

$$y = a_1 \# a_2 \# \cdots \# a_k$$

Since $a_i \leq \log_2 N < |x| + 1$, the binary encoding of w has length

$$\begin{aligned} |y| &= 2(|a_1| + \cdots + |a_k| + k - 1) \\ &< 2k(\log_2 |x| + 1) < |x| \end{aligned}$$

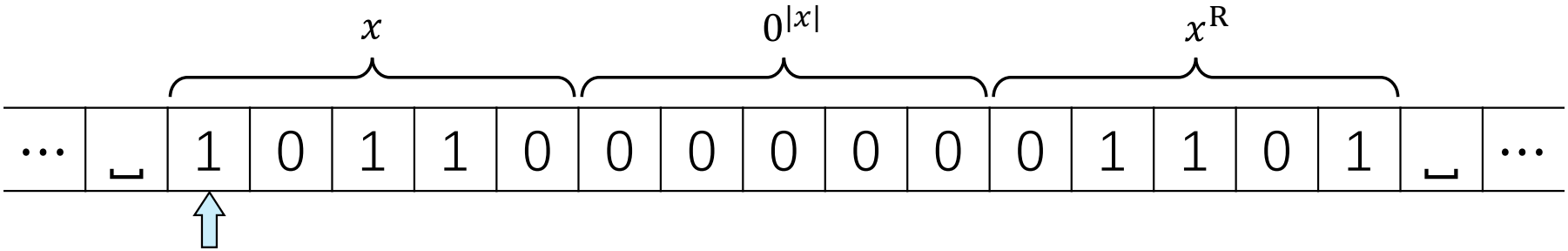
for every x with large enough length, contradicting the incompressibility.

Applications of Incompressibility

Example 2: Every single-tape, k-state Turing machine M that decides the language of palindromes $L = \{ w \in \{0,1\}^* \mid w = w^{\mathrm{R}} \}$ must take at least $\frac{1}{18k}|w|^2$ steps on some input w.

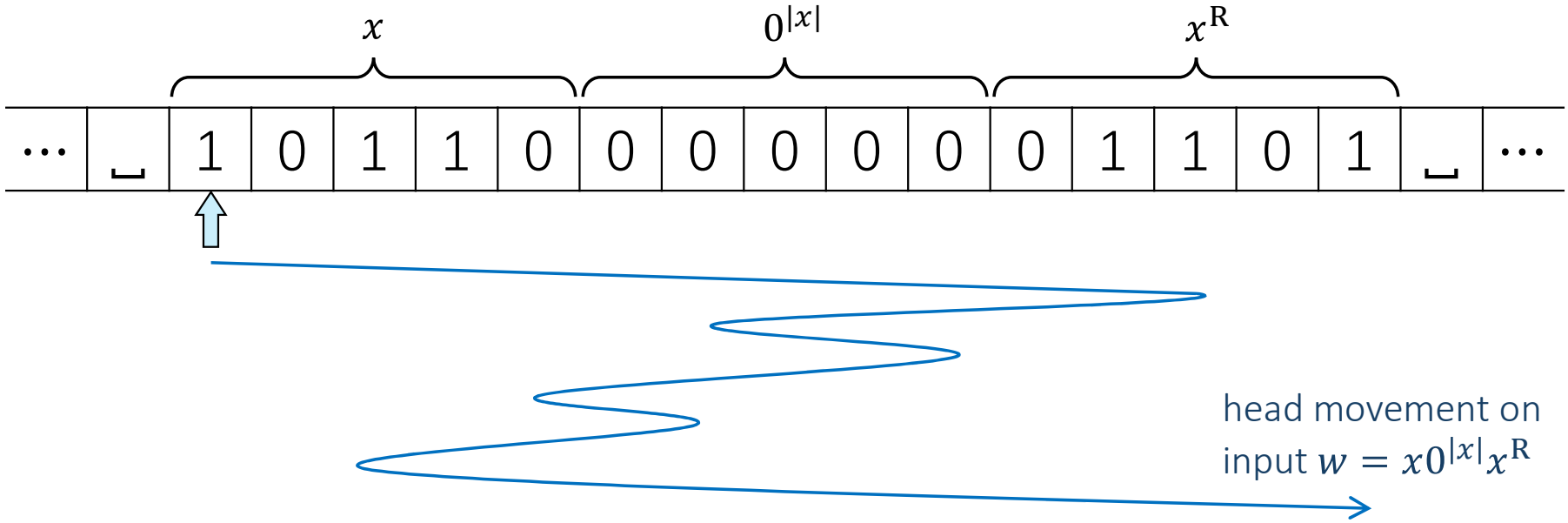
Applications of Incompressibility

Example 2: Every single-tape, k-state Turing machine M that decides the language of palindromes $L = \{ w \in \{0,1\}^* \mid w = w^{\mathrm{R}} \}$ must take at least $\frac{1}{18k}|w|^2$ steps on some input w.



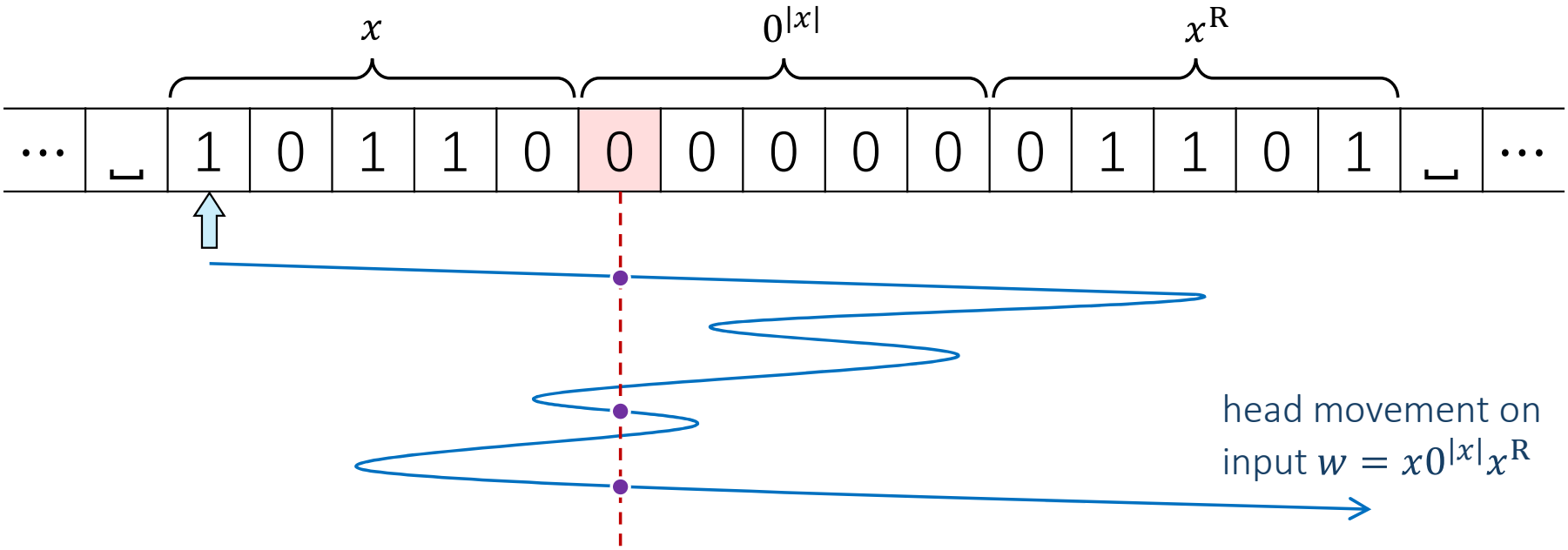
Applications of Incompressibility

Example 2: Every single-tape, k-state Turing machine M that decides the language of palindromes $L = \{ w \in \{0,1\}^* \mid w = w^{\mathrm{R}} \}$ must take at least $\frac{1}{18k}|w|^2$ steps on some input w.



Applications of Incompressibility

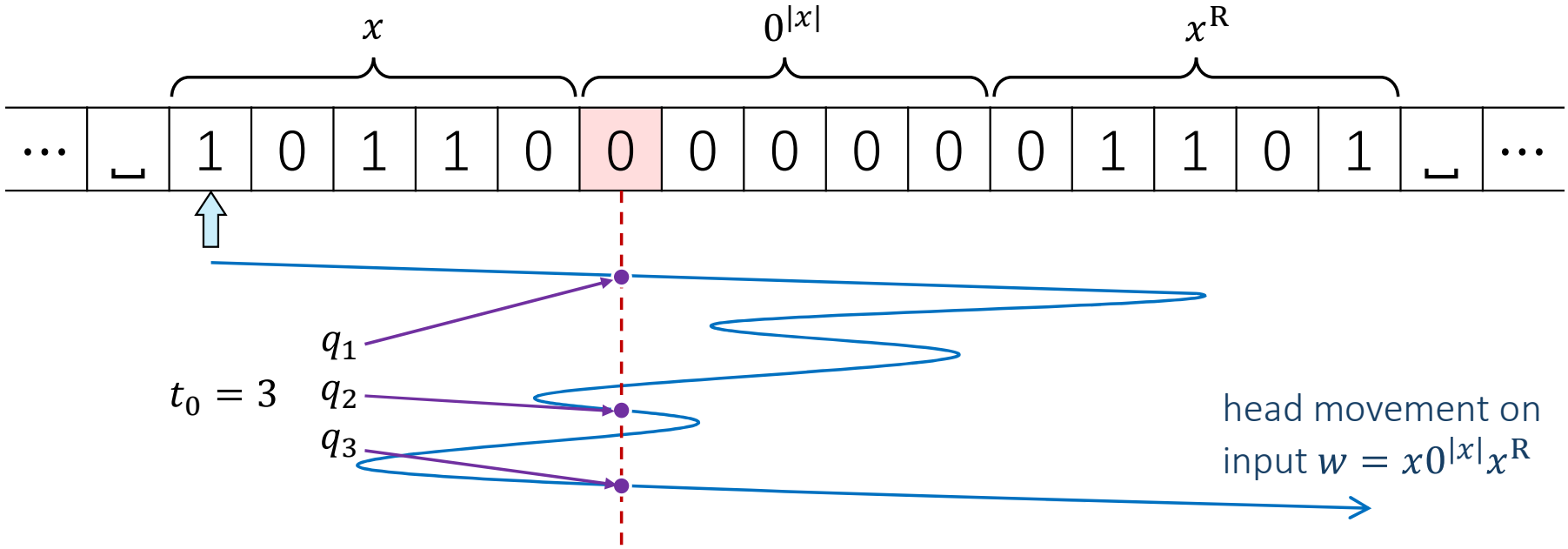
On input $w = x0^{|x|}x^{\mathrm{R}}$, consider the state at each time the head moves **right** onto the first cell in $0^{|x|}$.



Applications of Incompressibility

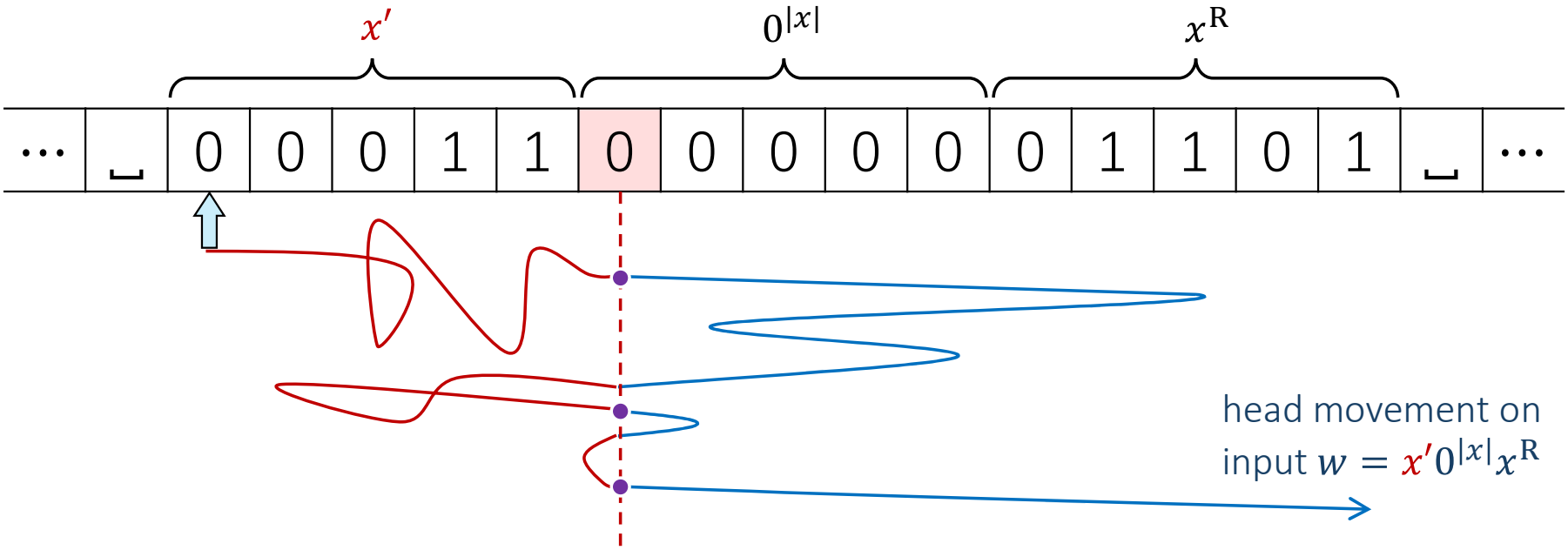
On input $w = x0^{|x|}x^{\mathrm{R}}$, consider the state at each time the head moves **right** onto the first cell in $0^{|x|}$. Describe x using these states:

$$y = |x| \# q_1 q_2 \cdots q_{t_0}$$



Applications of Incompressibility

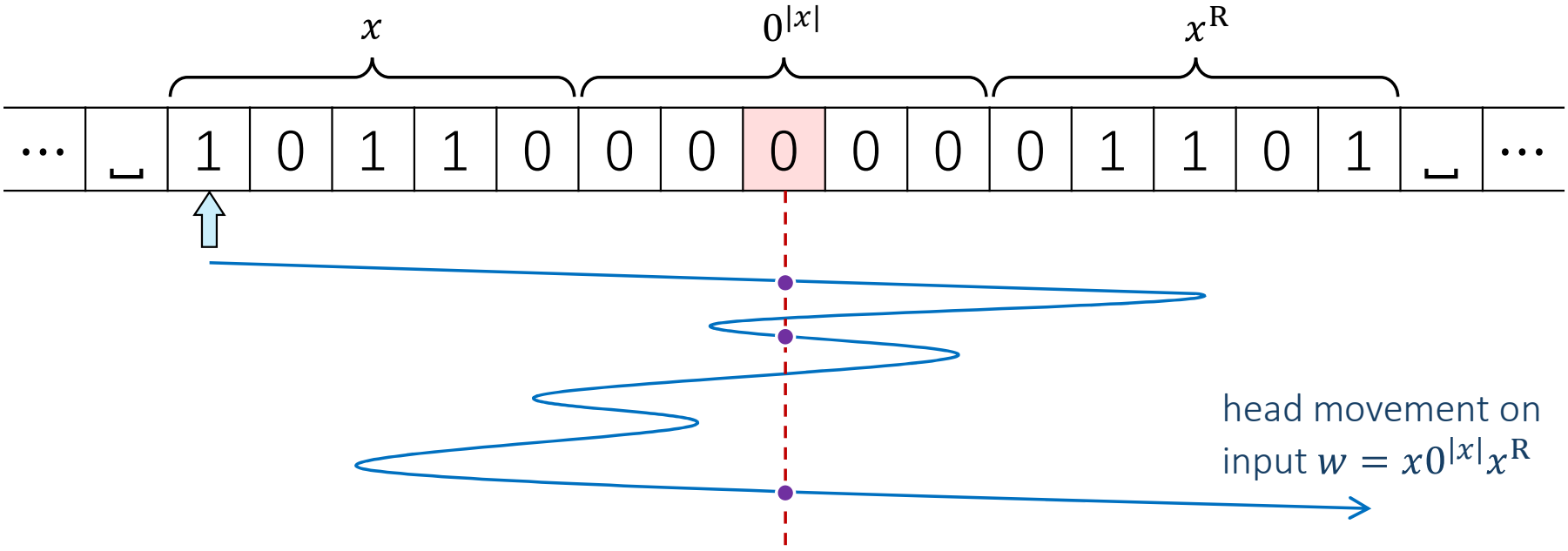
We can recover x from y, since if two different strings $x \neq x'$ shares the same description, then $|x| = |x'|$ and $x'0^{|x|}x^{\mathrm{R}}$ will be accepted.



Applications of Incompressibility

The same argument holds for every cell in the middle block $0^{|x|}$, so we take the i-th one with the smallest t_i and use description:

$$y = |x| \# i \# q_1 q_2 \cdots q_{t_i}$$



Applications of Incompressibility

The same argument holds for every cell in the middle block $0^{|x|}$, so we take the i-th one with the smallest t_i and use description:

$$y = |x| \# i \# q_1 q_2 \cdots q_{t_i}$$

If M always runs in $\frac{1}{18}|w|^2 = \frac{1}{2k}|x|^2$ steps, then $t_i \leq \frac{1}{2k}|x|$, and thus the binary encoding of y has length

$$|y| \leq (2\log_2|x| + t_i + 2) \cdot (\log_2 k + 1)$$

$$\leq (2\log_2|x| + |x|/(2k) + 2) \cdot k < |x|$$

for every x with large enough length, contradicting the incompressibility.