Compression by deterministic pushdown automata

Olivier Carton[1*] Sylvain Perifel[1]

2022 – ANR Delta

[1]IRIF – Université Paris Cité & CNRS
[*]Supported by LIA SINFIN

Outline

Normality

Characterization by non-compressibility

Non-deterministic pushdown transducers

Deterministic pushdown transducers

Outline

Normality (Borel 1909)

The number of occurrences of a word w in a word u is

$$\mathrm{occ}(u, w) = \#\{i : u[i{:}i + |w| - 1] = w\}$$

A sequence $x \in A^{\mathbb{N}}$ normal if for each $w \in A^*$,

$$\lim_{n\to\infty} \frac{\mathrm{occ}(x[1{:}n], w}{n} = \frac{1}{(\#A)^{|w|}}.$$

For $A = \{0, 1\}$, this means

- the frequencies in x of the 2 digits 0 and 1 are $1/2$,
- the frequencies in x of the 4 words $00, 01, 10, 11$ are $1/4$,
- the frequencies in x of the 8 words $000, 001, \ldots, 111$ are $1/8$,
- $\ldots$

Examples

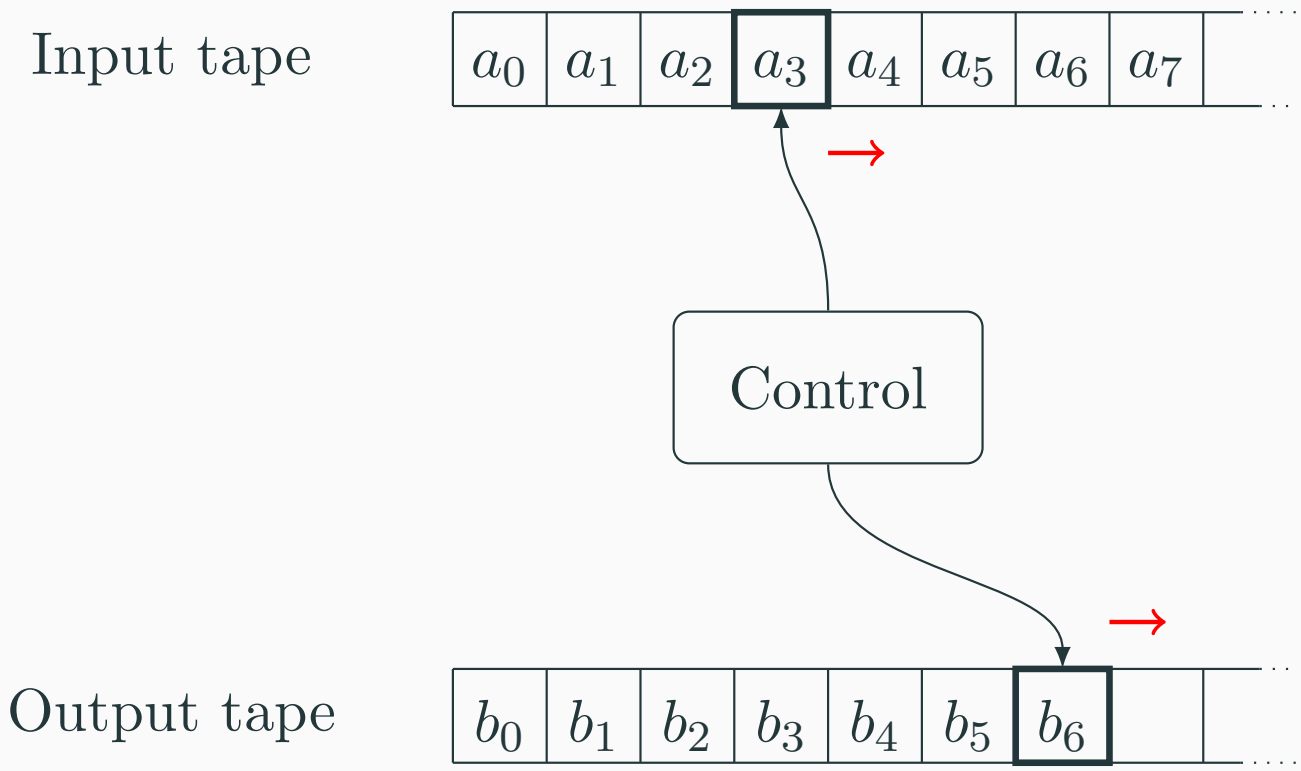
Theorem (Borel 1909)
Almost all real numbers are normal, that is, the measure of the set of normal numbers in $[0, 1)$ *is* 1.

Examples

- the Champernowne sequence $01\,00011011\,000001010\cdots$,
- the sequence of primes in binary $11011101111110111101\cdots$.
- It is not known whether all digits occur infinitely many times in the base 3 expansion of $\sqrt{2}$.

Outline

Transducers



Transducers as compressors

A transducer $\mathcal{T}$ *compresses* a sequence $x = a_1 a_2 a_3 \cdots$ if

- the function (relation) realized by $\mathcal{T}$ is one-to-one,
- there is an accepting run with input x

$$C_0 \xrightarrow{a_1|v_1} C_1 \xrightarrow{a_2|v_2} C_2 \xrightarrow{a_3|v_3} C_3 \cdots$$

such that

$$\liminf_{n\to\infty} \frac{|v_1 v_2 \cdots v_n| \log \#B}{|a_1 a_2 \cdots a_n| \log \#A} < 1.$$

Characterization of normal sequences

Theorem (Many people)
A sequence is normal if and only if it cannot be compressed by deterministic one-to-one finite state transducer.

- Schnorr and Stimm (1971)
  non-normality $\Leftrightarrow$ finite-state martingale success
- Dai, Lathrop, Lutz and Mayordomo (2004)
  compressibility $\Leftrightarrow$ finite-state martingale success
  normality $\Rightarrow$ no martingale success
- Bourke, Hitchcock and Vinodchandran (2005)
  non-normality $\Rightarrow$ martingale success
- Becher and Heiber (2013)
  non-normality $\Leftrightarrow$ compressibility (direct)

Summary of the known results

	det	non-det	non-rt
finite-state	N	N	N
1 counter	N	N	N
$\geqslant 2$ counters	N	N	T
1 stack	?	C	C
1 stack + 1 counter	C	C	T

where

N means *cannot compress normal sequences*

C means *can compress some normal sequence*

T means *is Turing complete* and thus can compress.

Outline

Non-deterministic pushdown transducers

Let A be an alphabet of even cardinality. Let w_n be the concatenation in some order of all words of length n.

$$x = w_1 \tilde{w}_1 w_2 \tilde{w}_2 w_3 \tilde{w}_3 \cdots$$

where $\tilde{w}$ is the reverse of w. For $A = \{0, 1\}$, it gives

$$x = 0\,1\,1\,0\,00\,01\,10\,11\,11\,01\,10\,00\,000\,001\,010\,\cdots$$

The run is an alternation of pushing and popping phases.

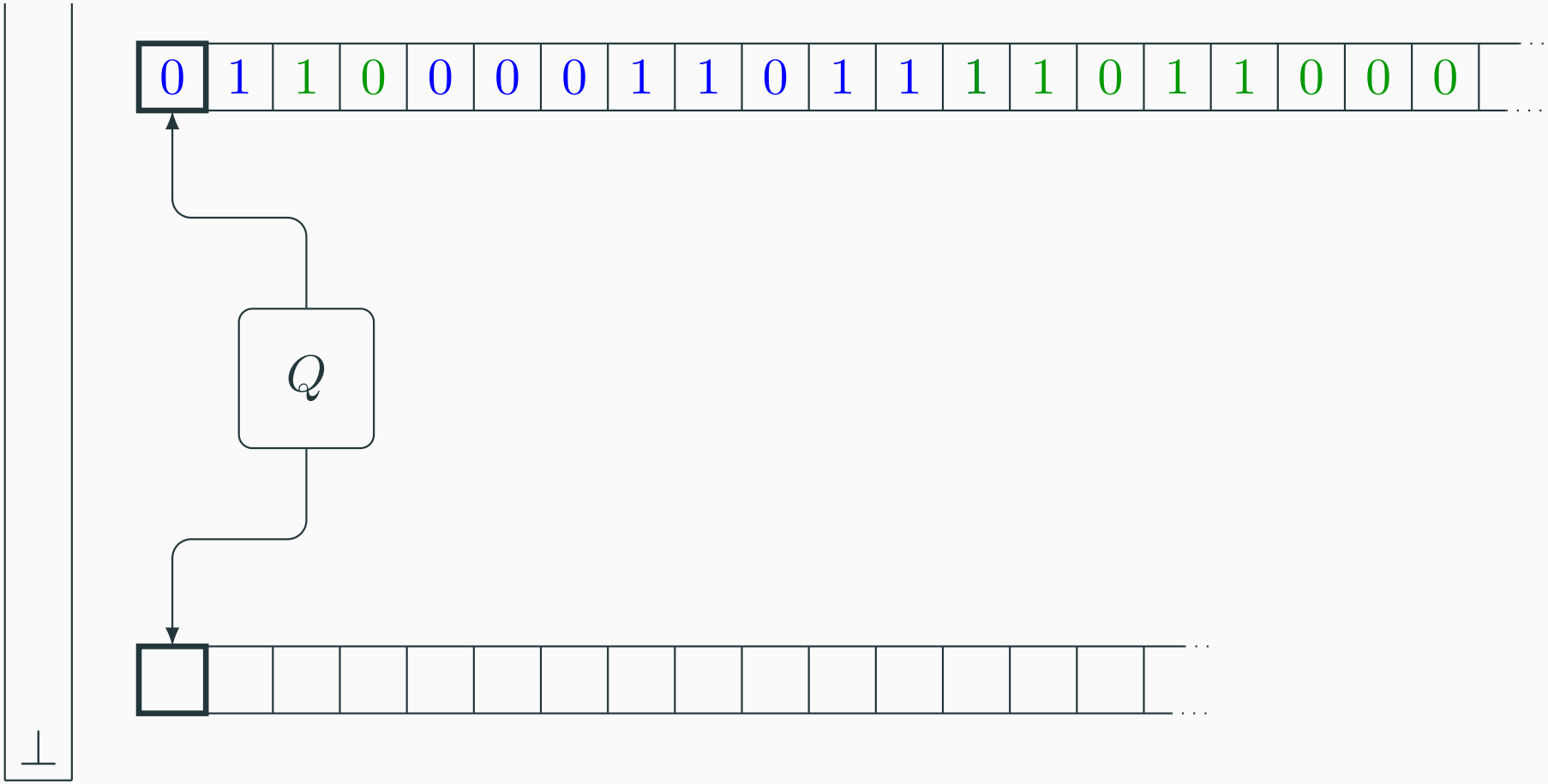
- In the pushing phase, each input symbol is pushed and output.
- In the popping phase, each input symbol must be equal to the stack top symbol and this top symbol is popped.
  A symbol $\square$ is output every two popped symbols.

Non-deterministic pushdown transducers

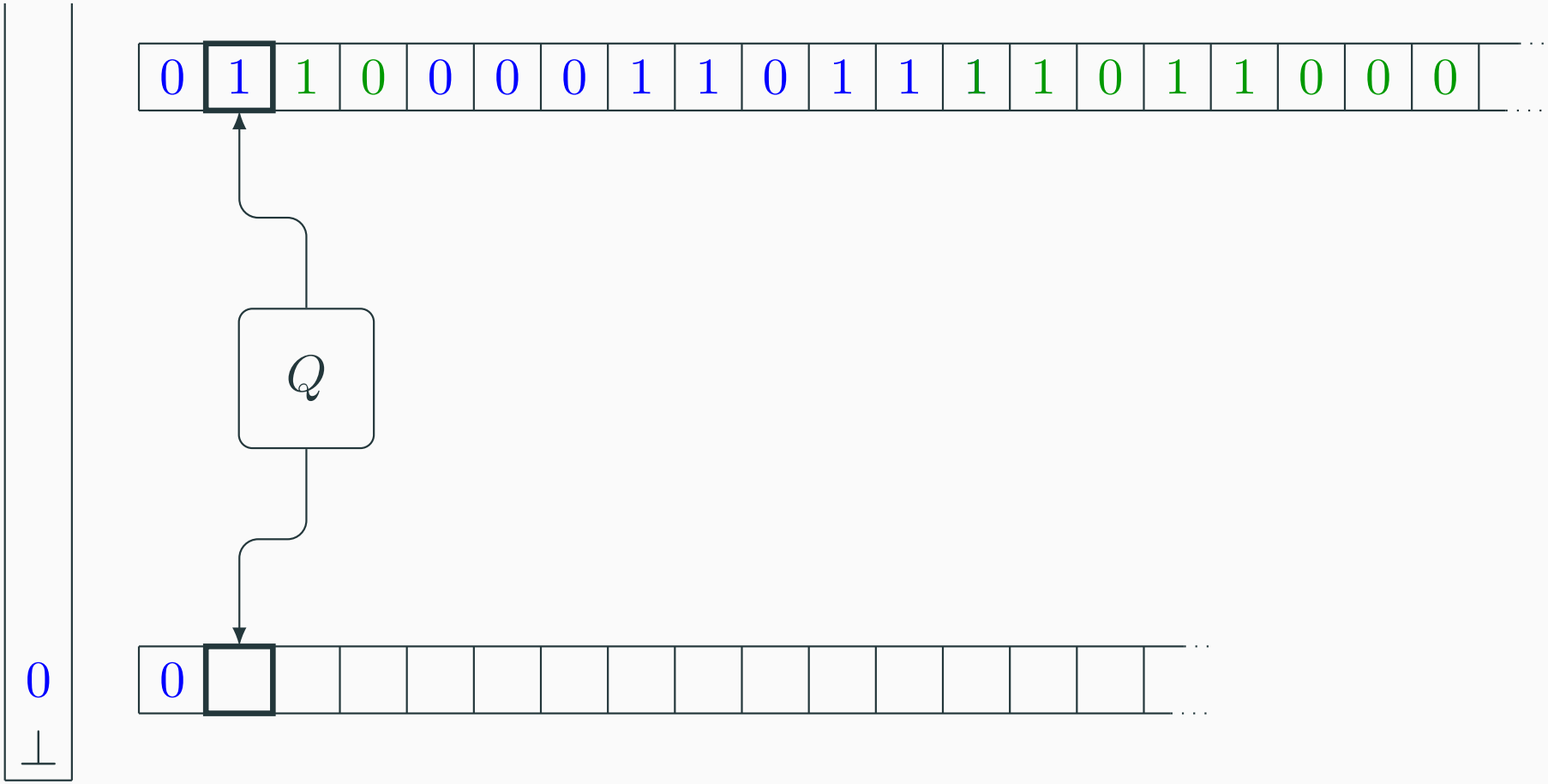
$$x = w_1\tilde{w}_1w_2\tilde{w}_2w_3\tilde{w}_3\cdots$$

- Each word w_n is read during a pushing phase.
- At the end of w_n, the transducer switches non-deterministically to a popping phase to read $\tilde{w}_n$.
- Each word $\tilde{w}_n$ is read during a popping phase.
- At the end of $\tilde{w}_n$, the stack is empty and the transducer switches back to a pushing phase to read w_{n+1}.
- While reading $w_n\tilde{w}_n$, the transducer outputs $w_n\square^{|w_n|/2}$.

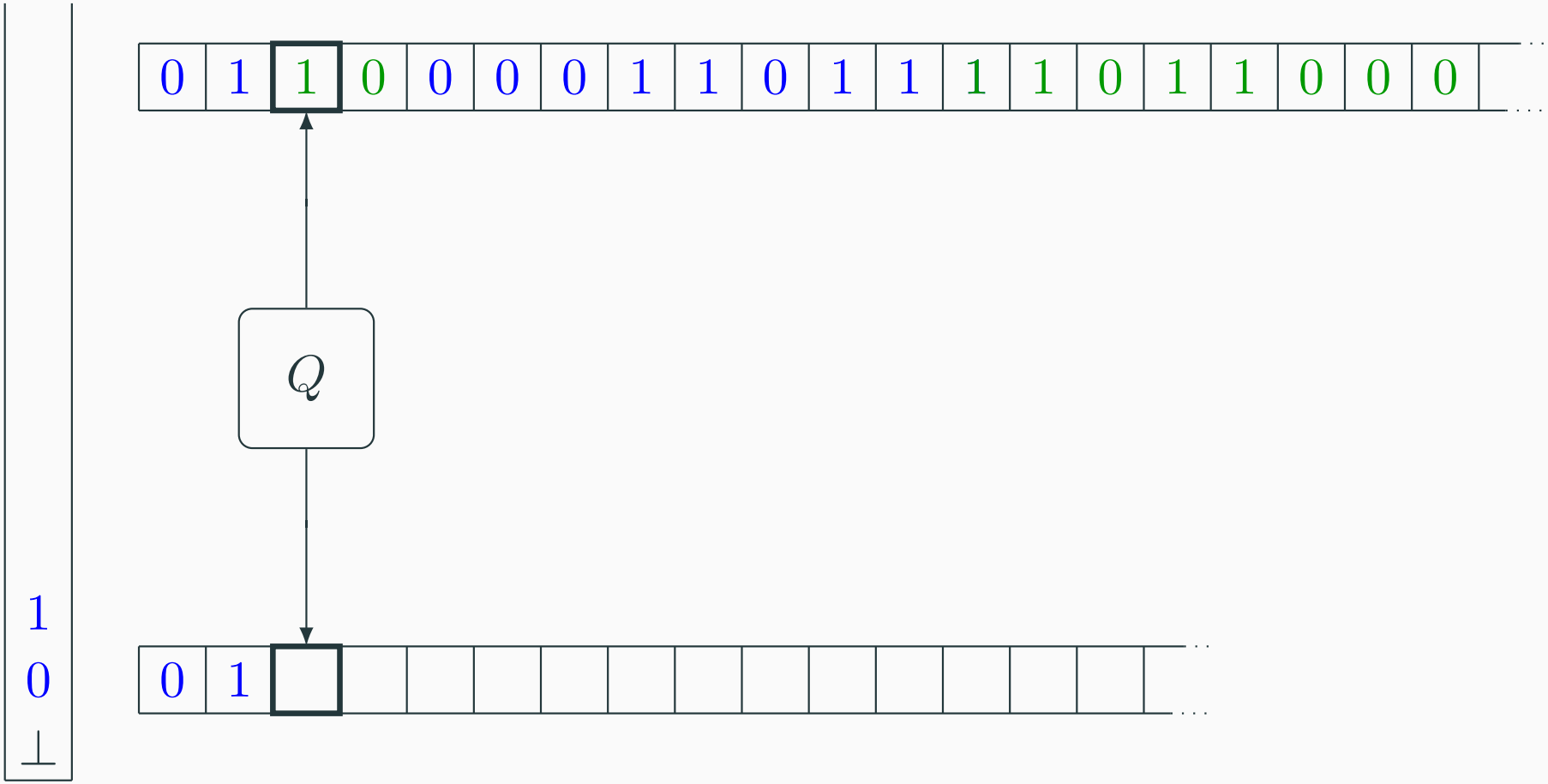
Let it run



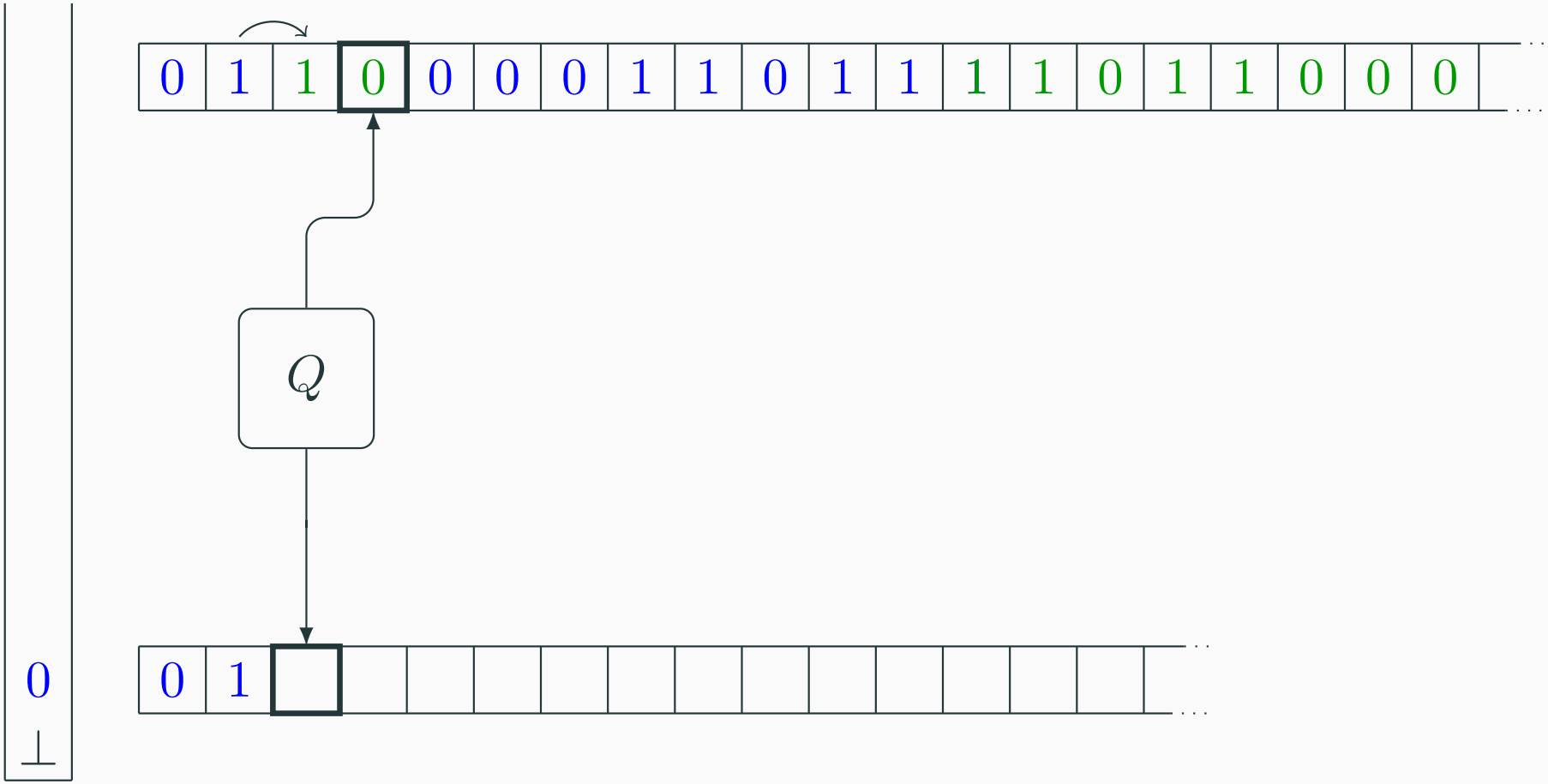
Let it run



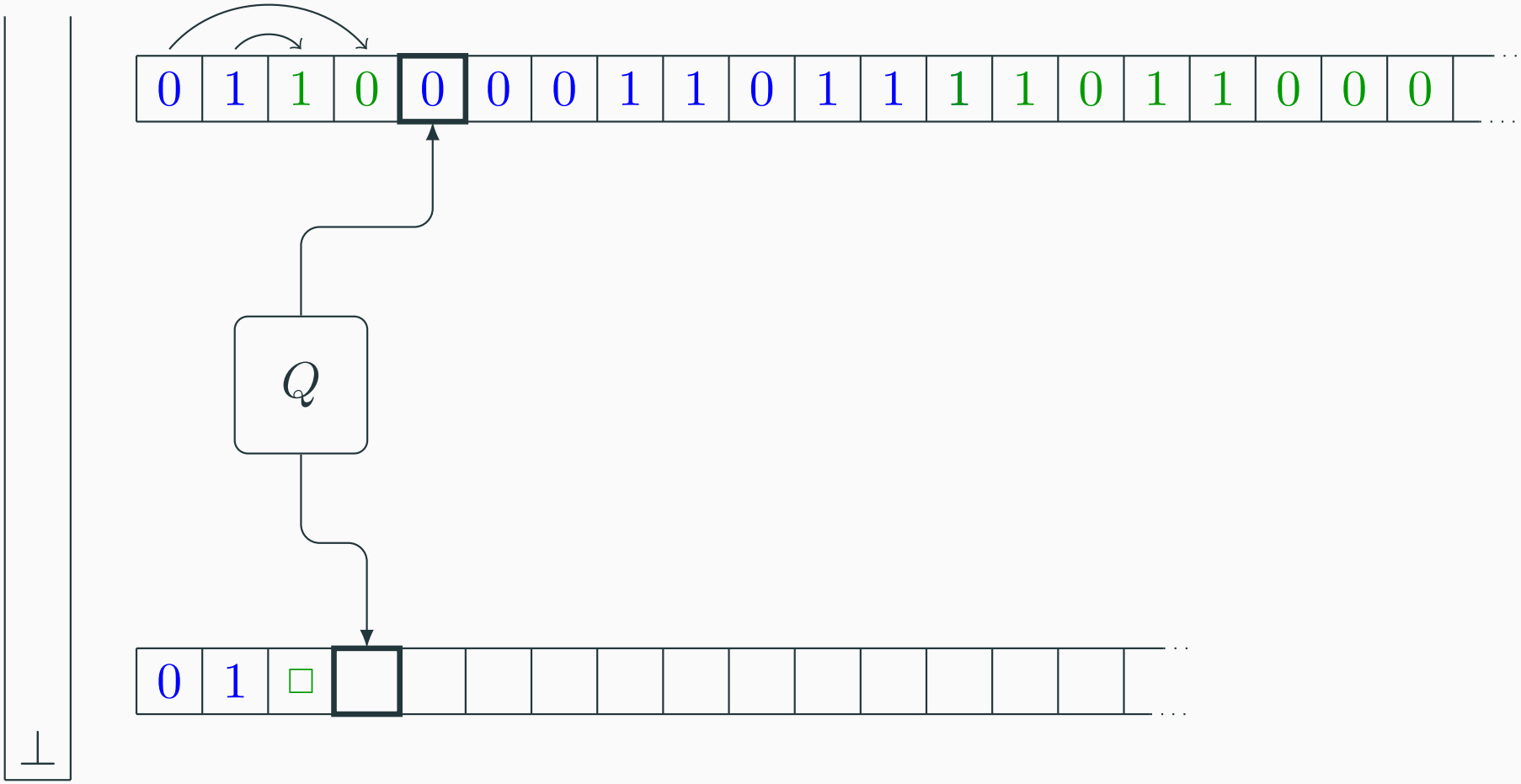
Let it run



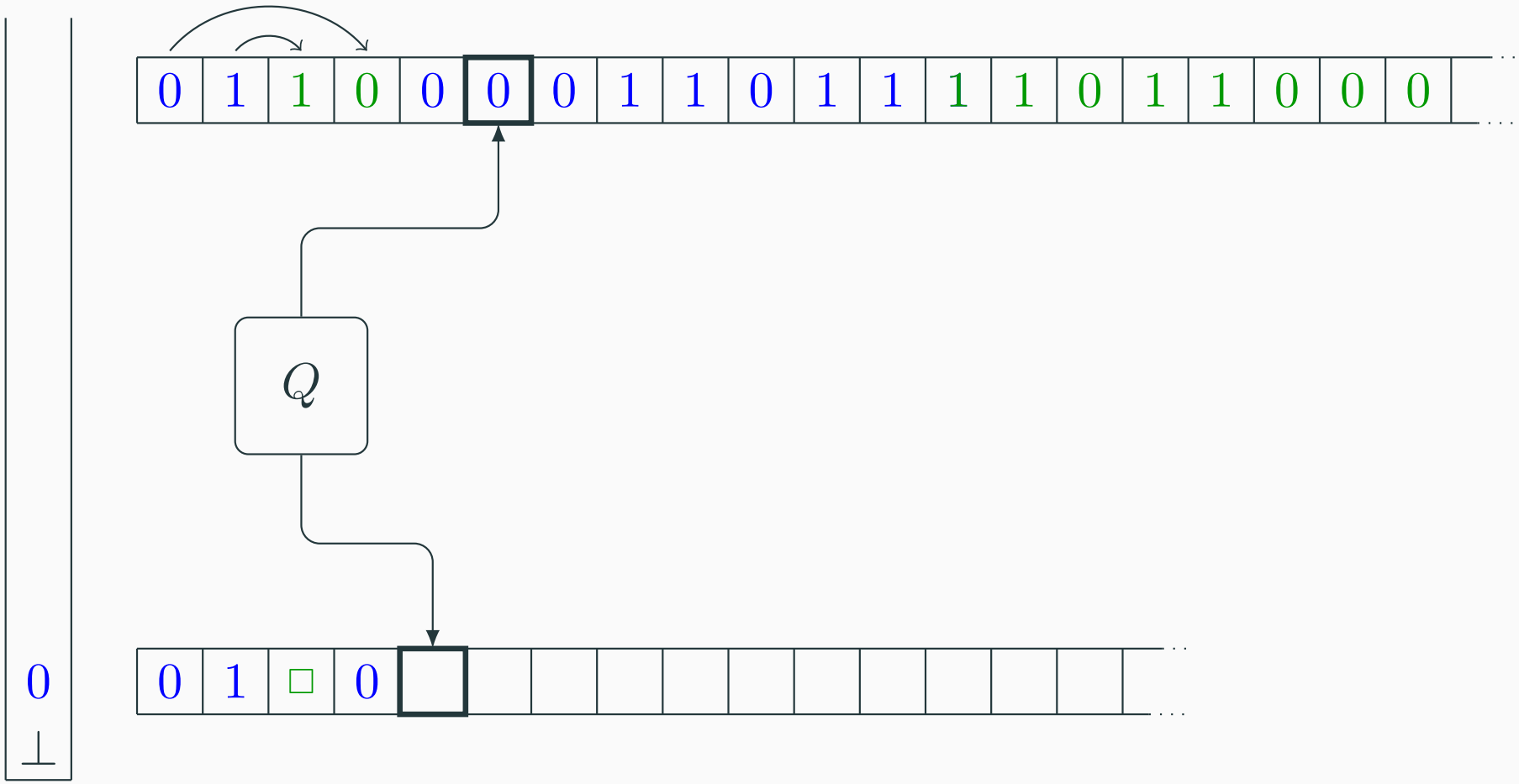
Let it run



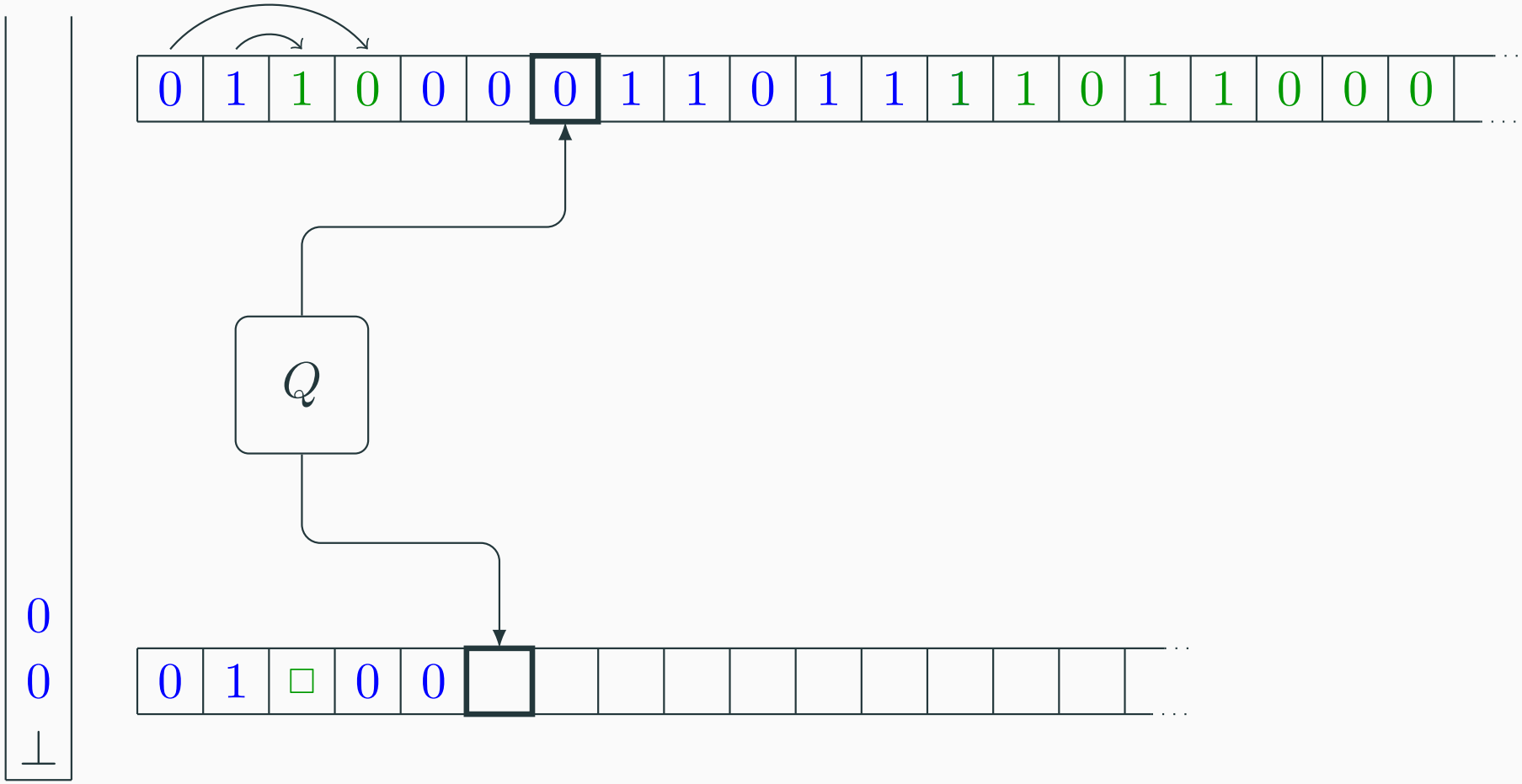
Let it run



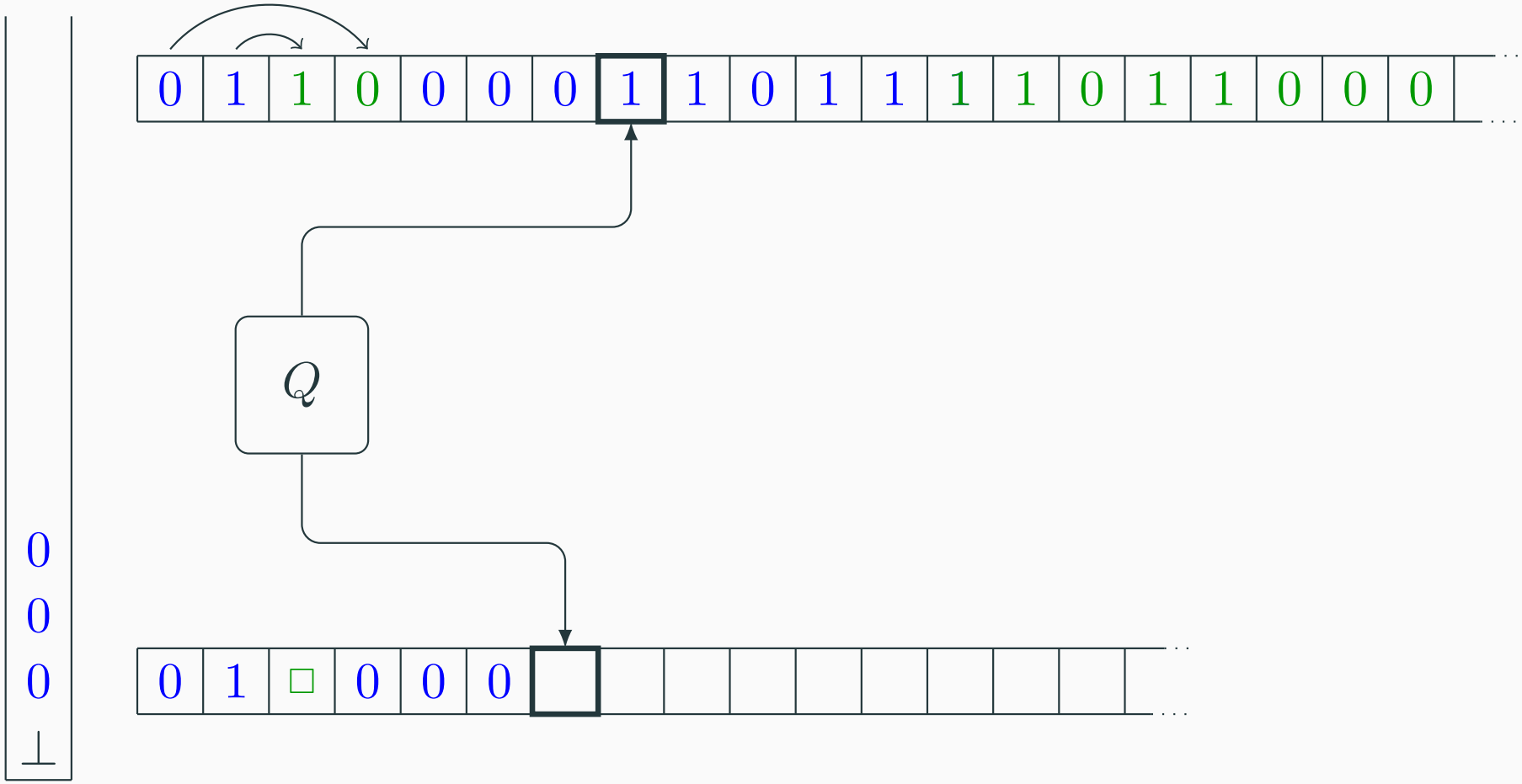
Let it run



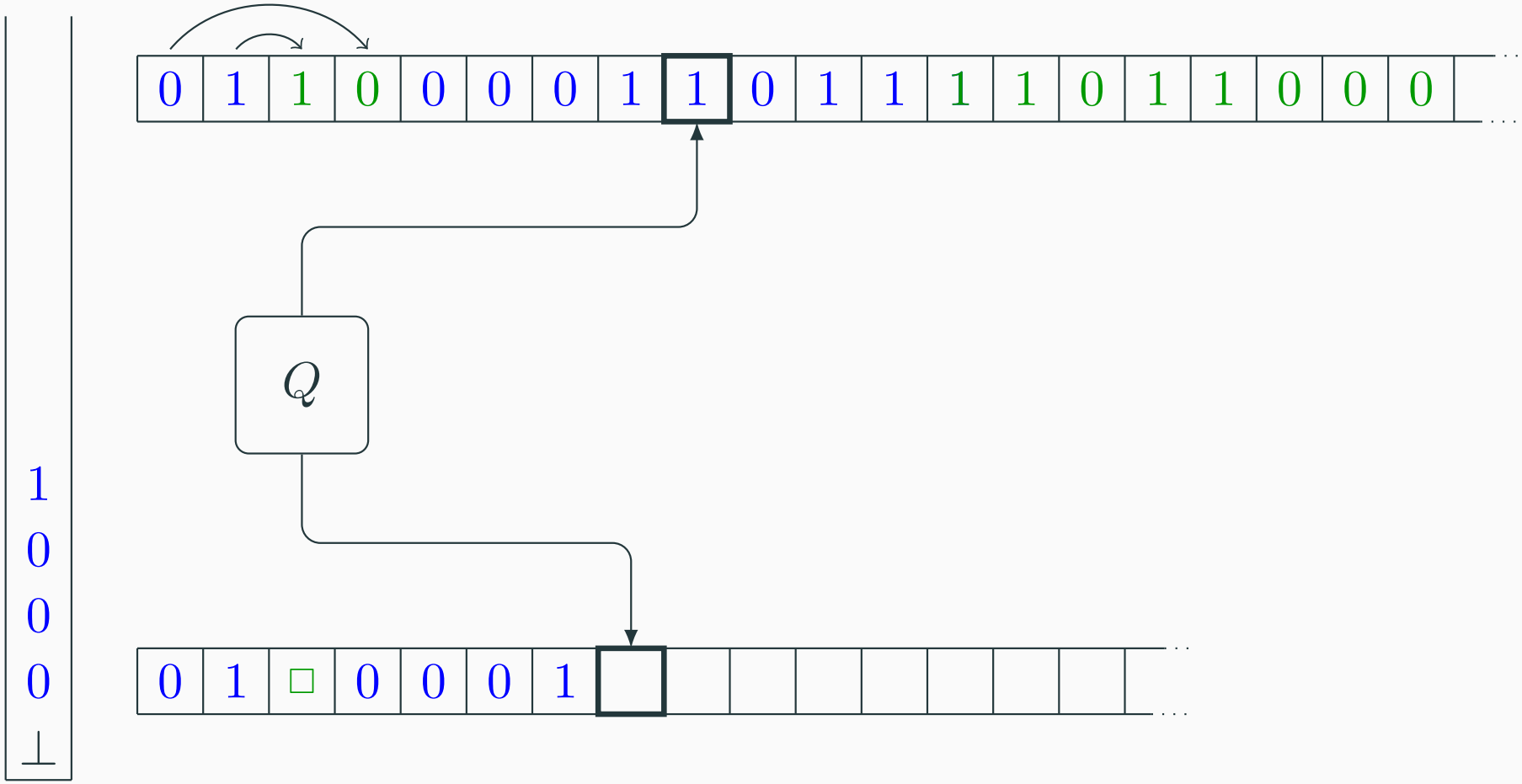
Let it run



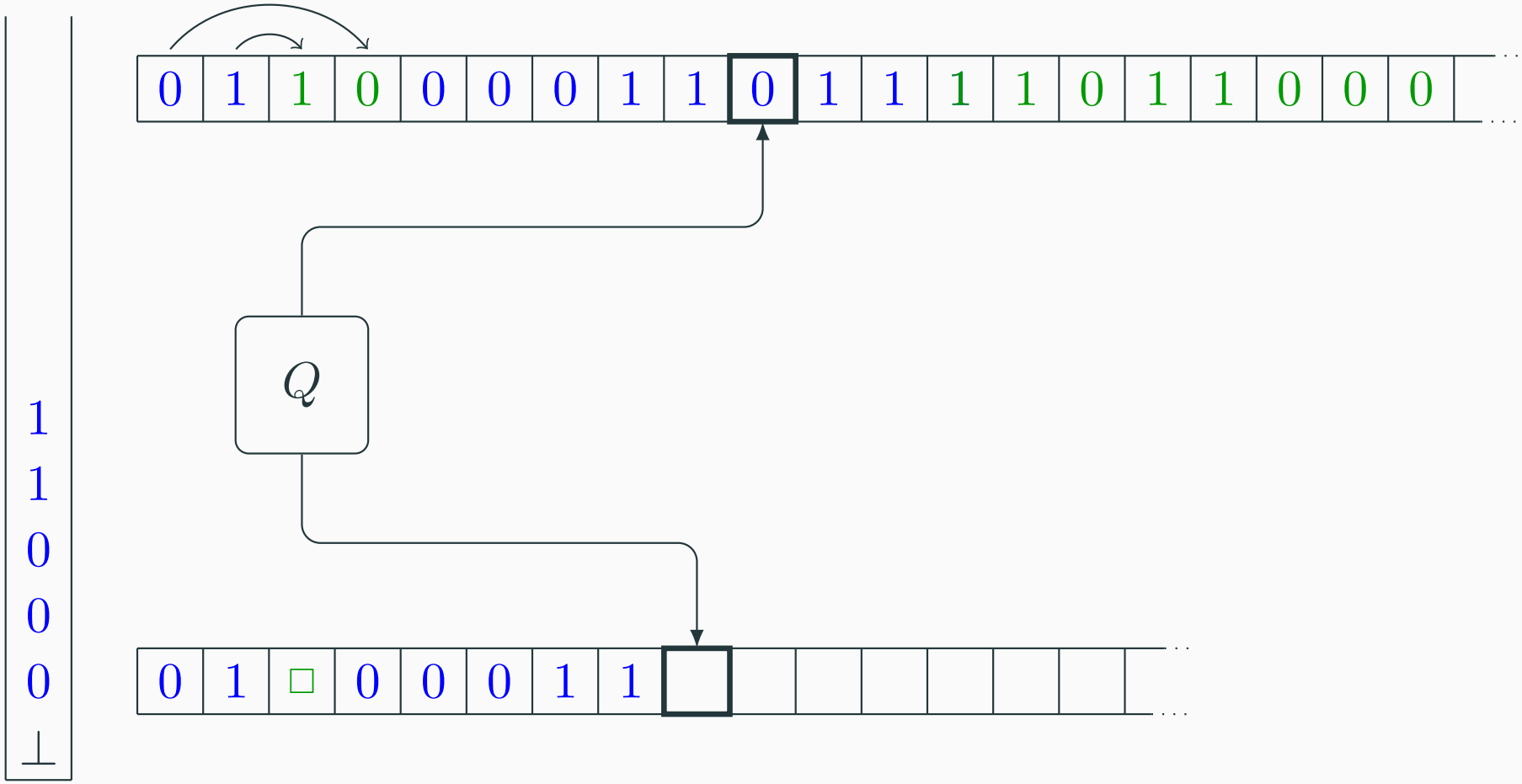
Let it run



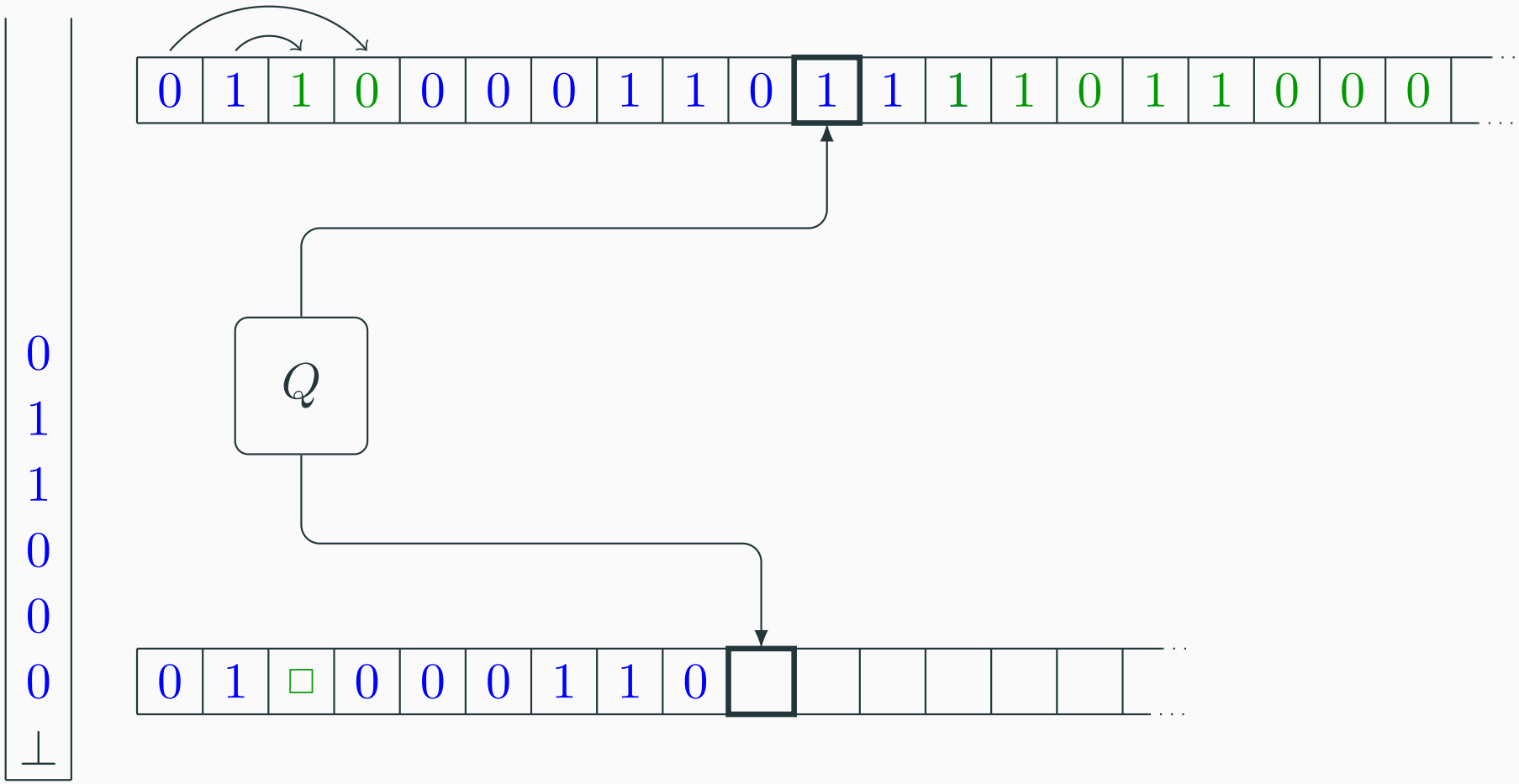
Let it run



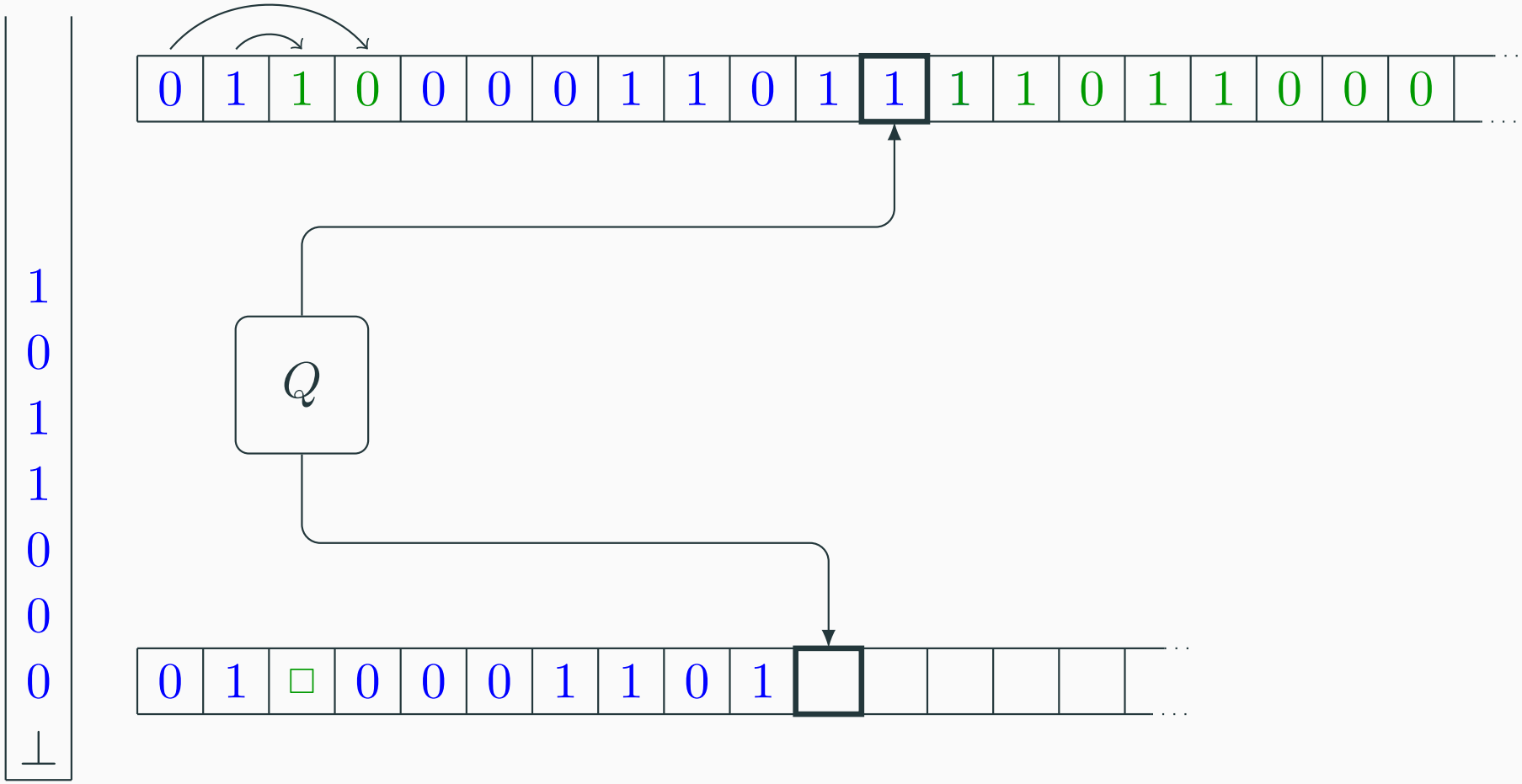
Let it run



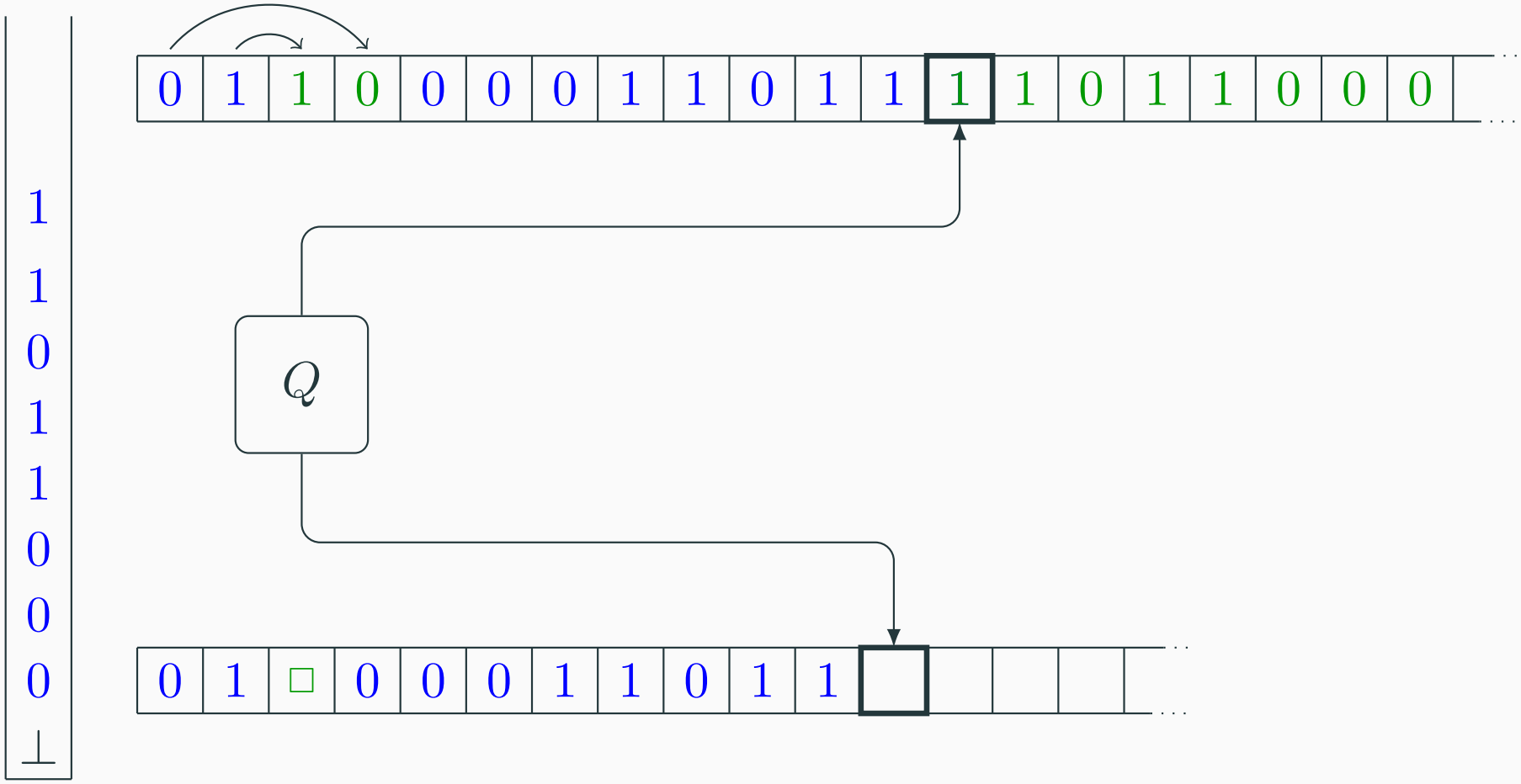
Let it run



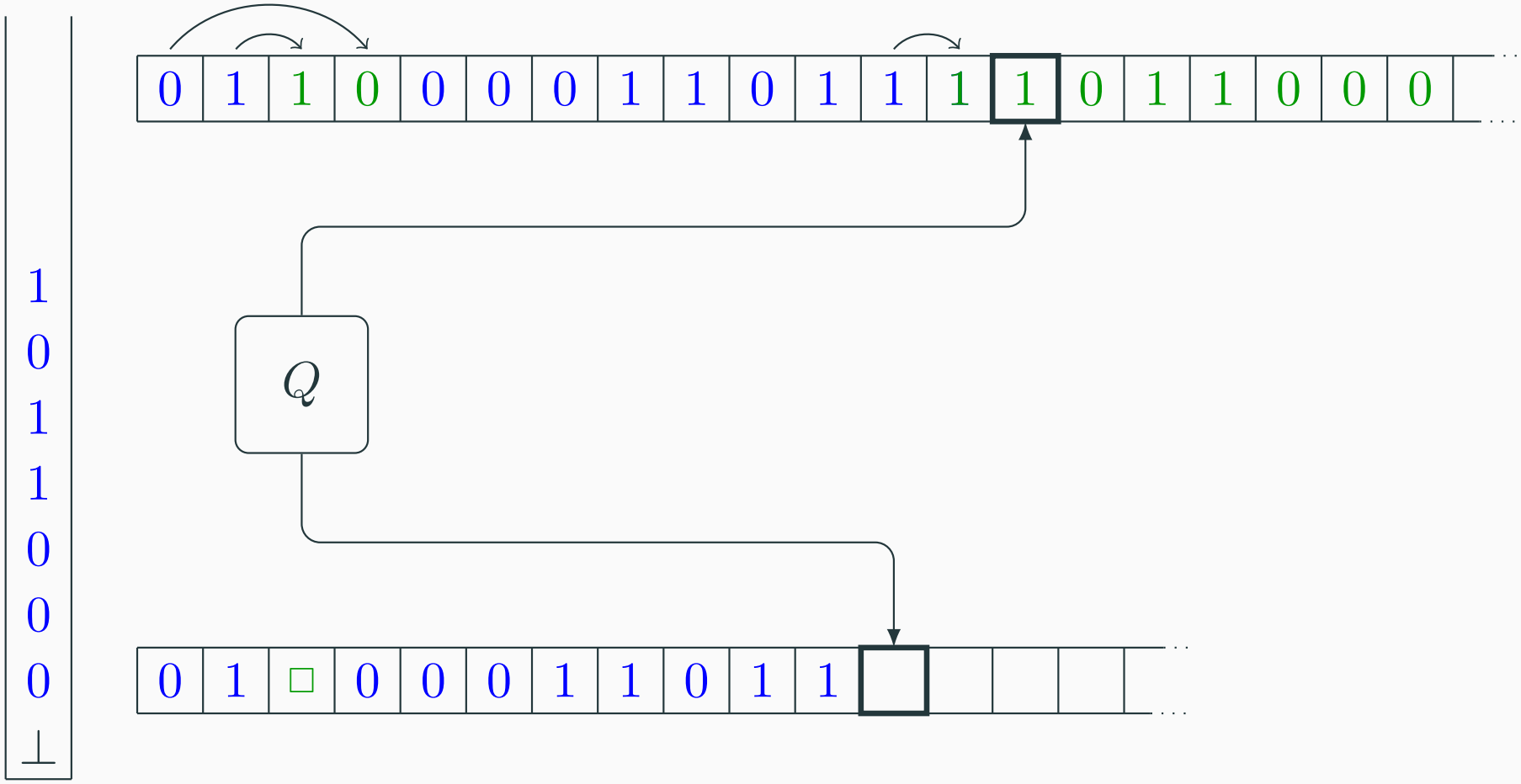
Let it run



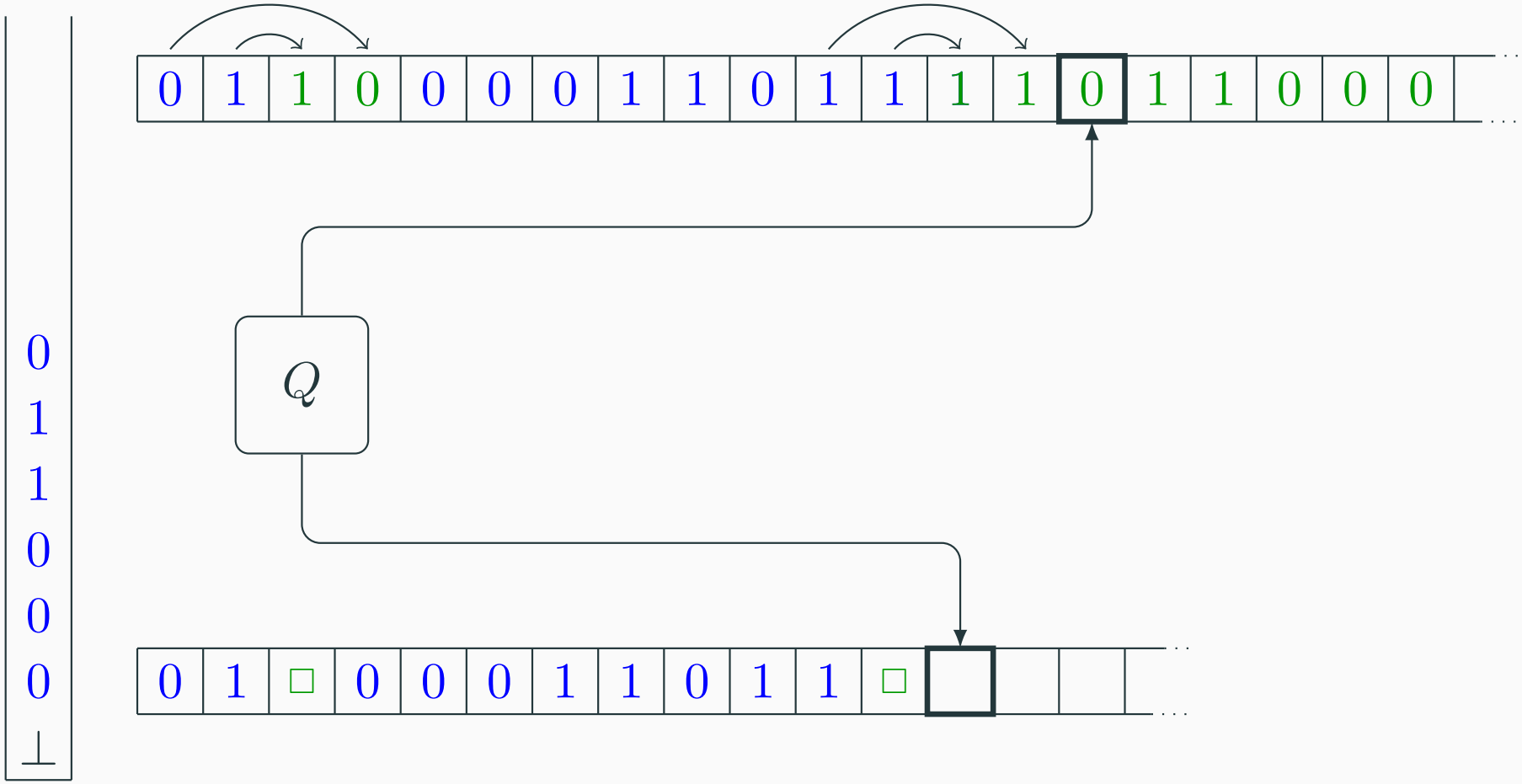
Let it run



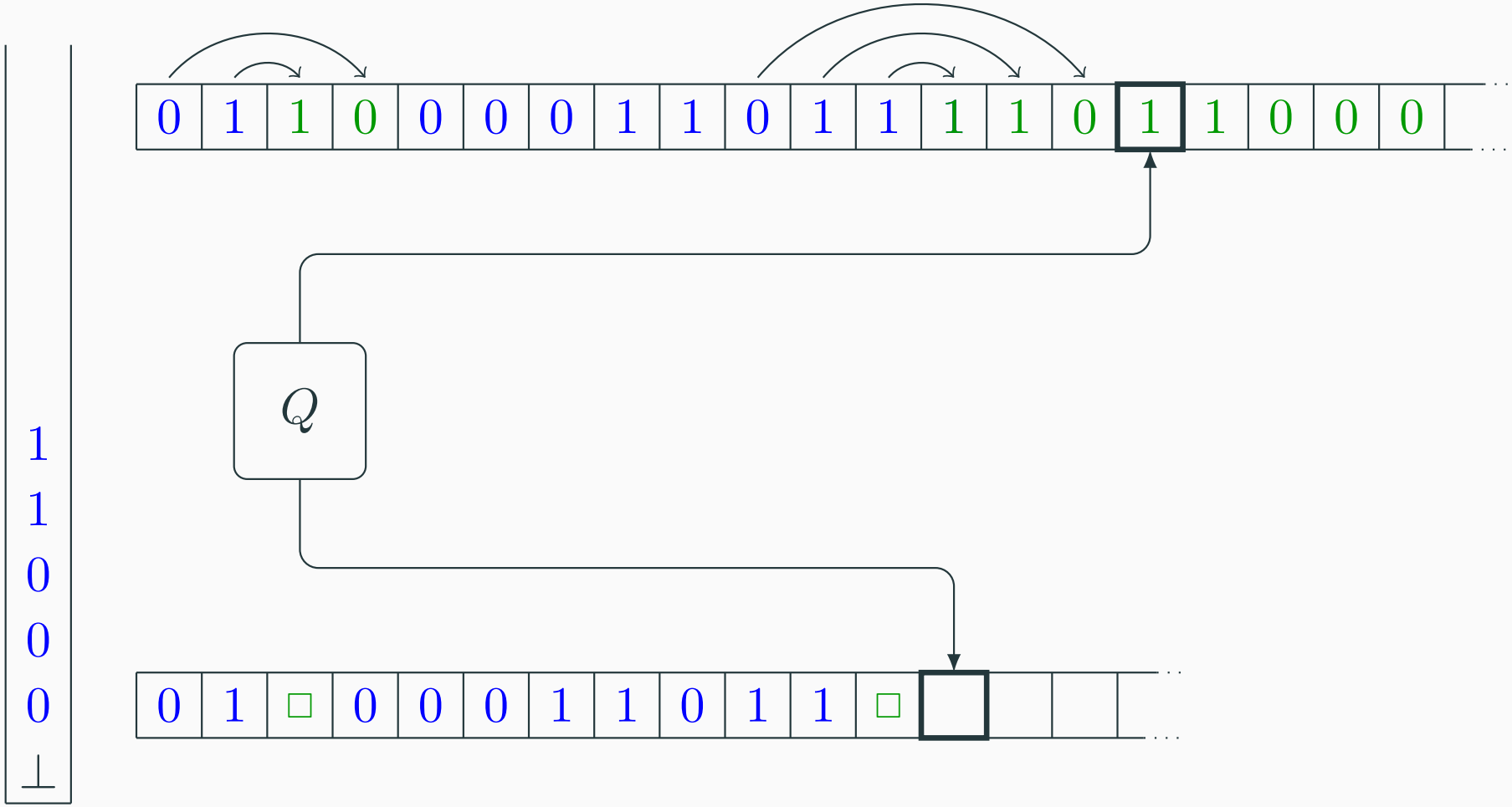
Let it run



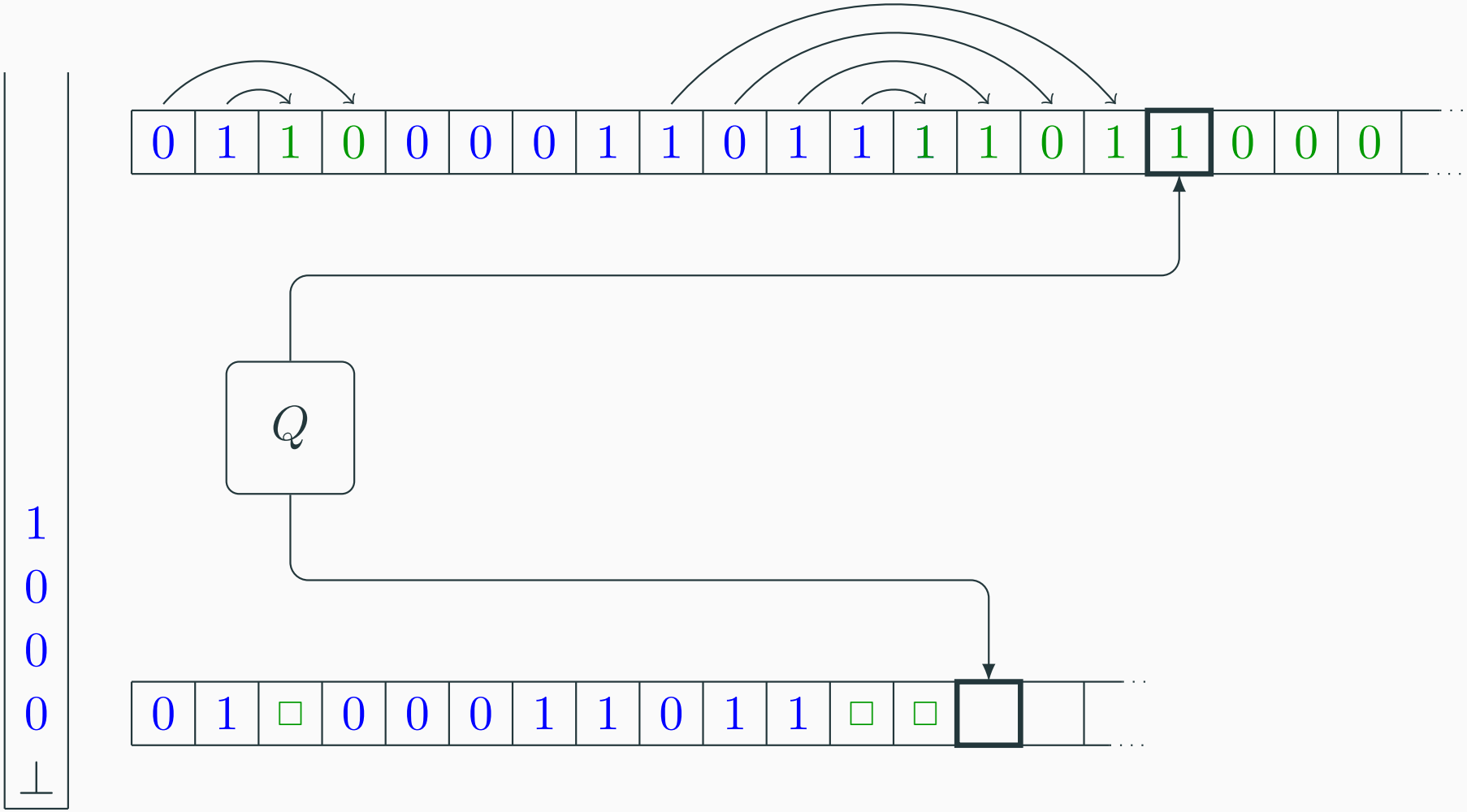
Let it run



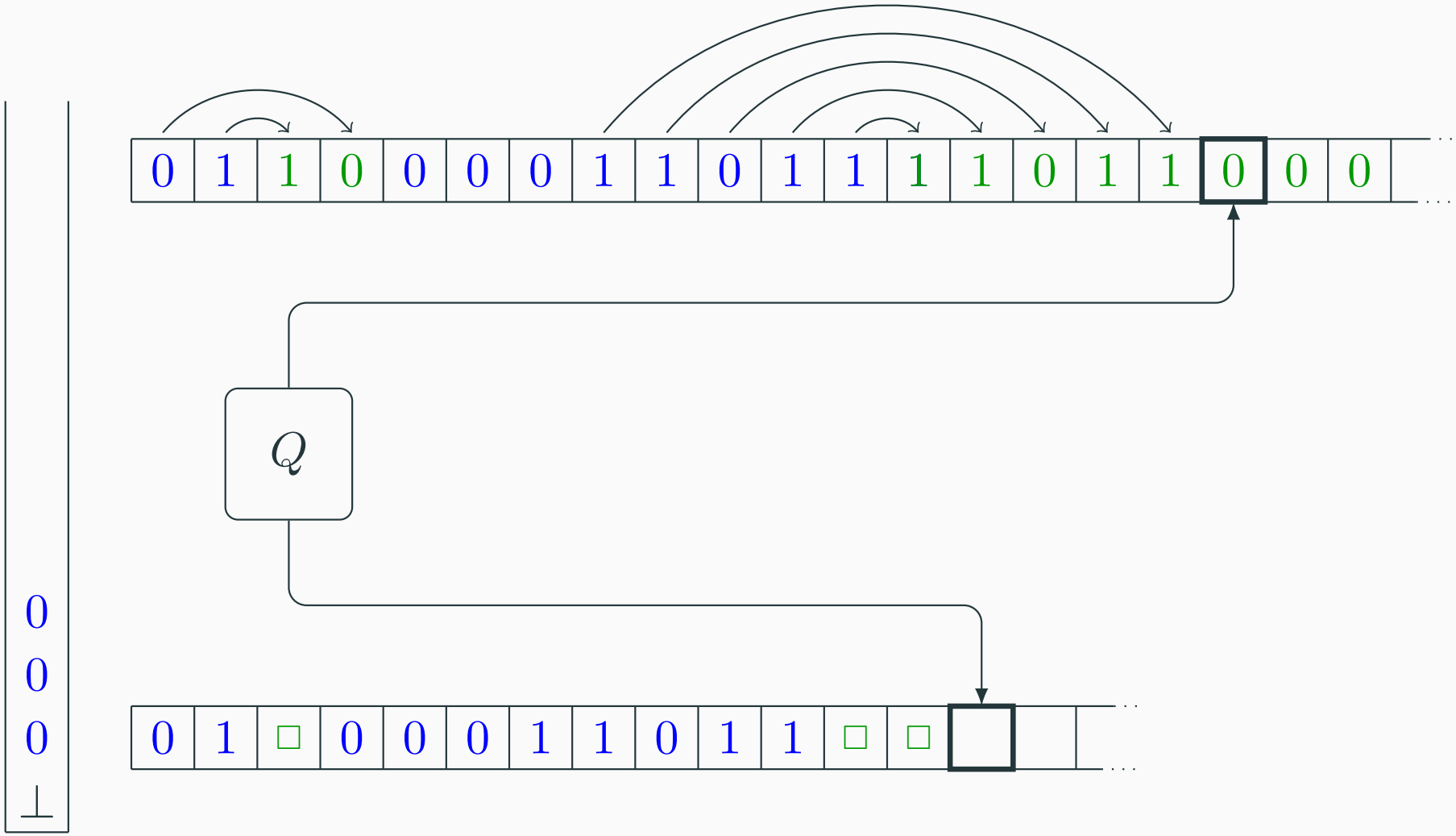
Let it run



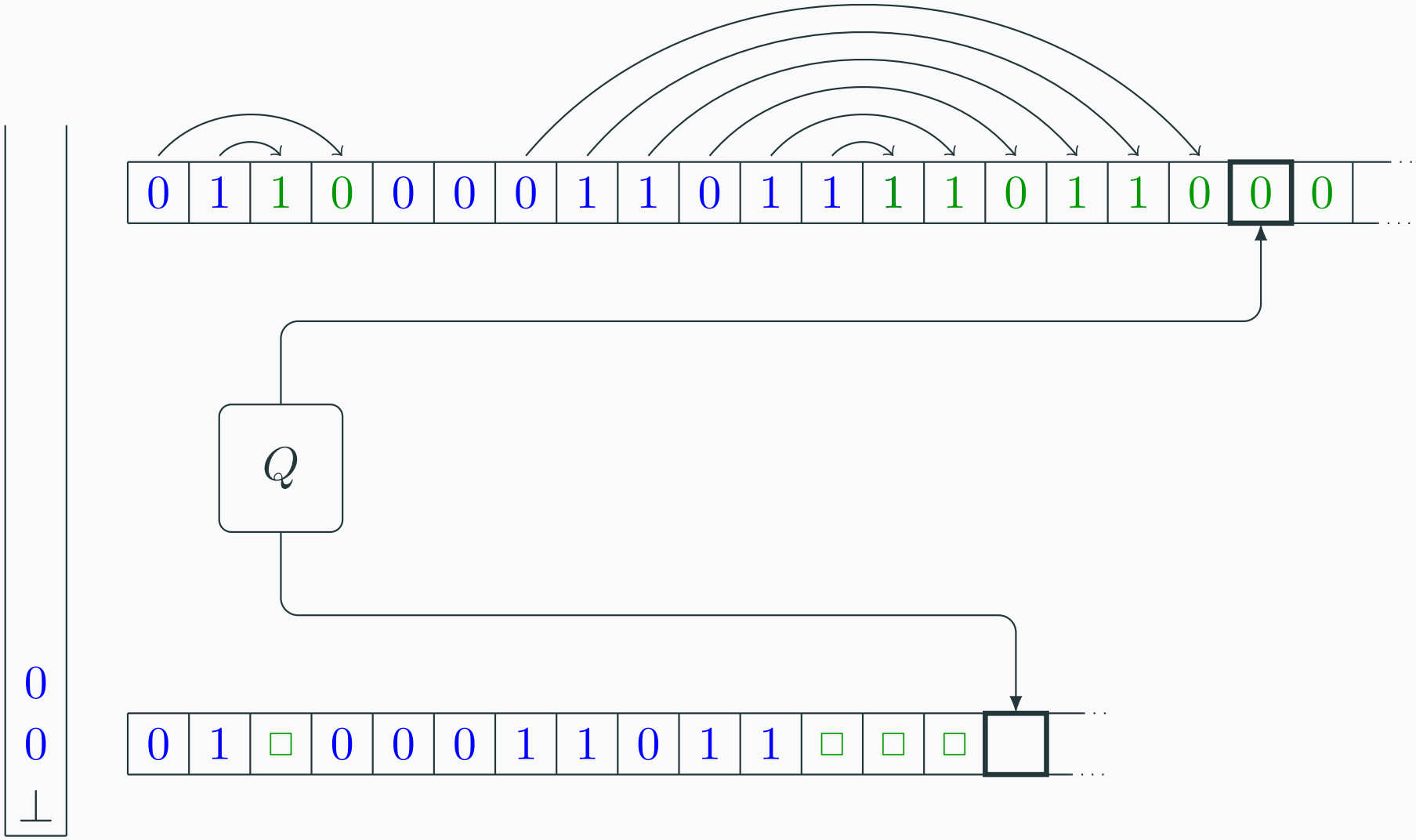
Let it run



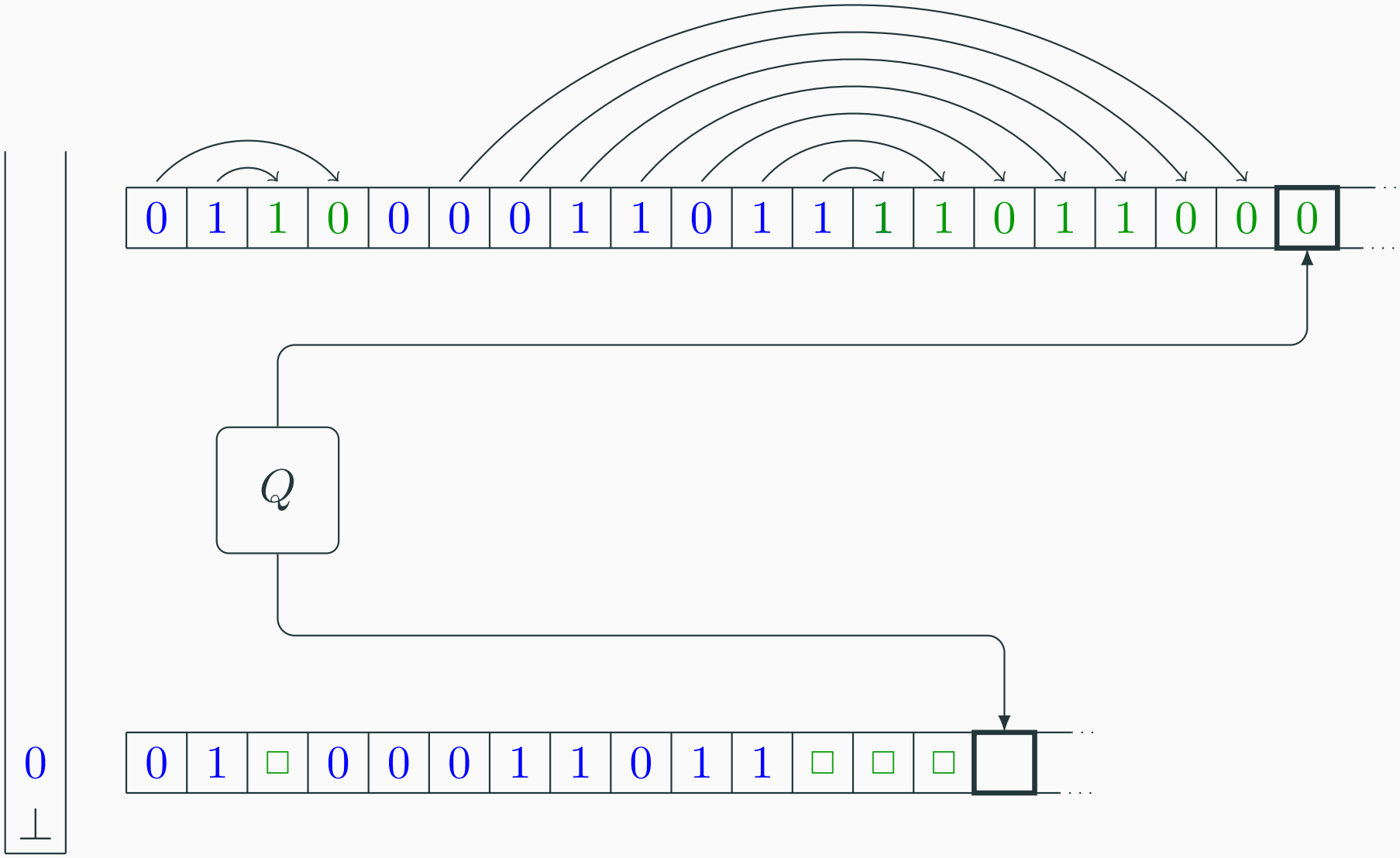
Let it run



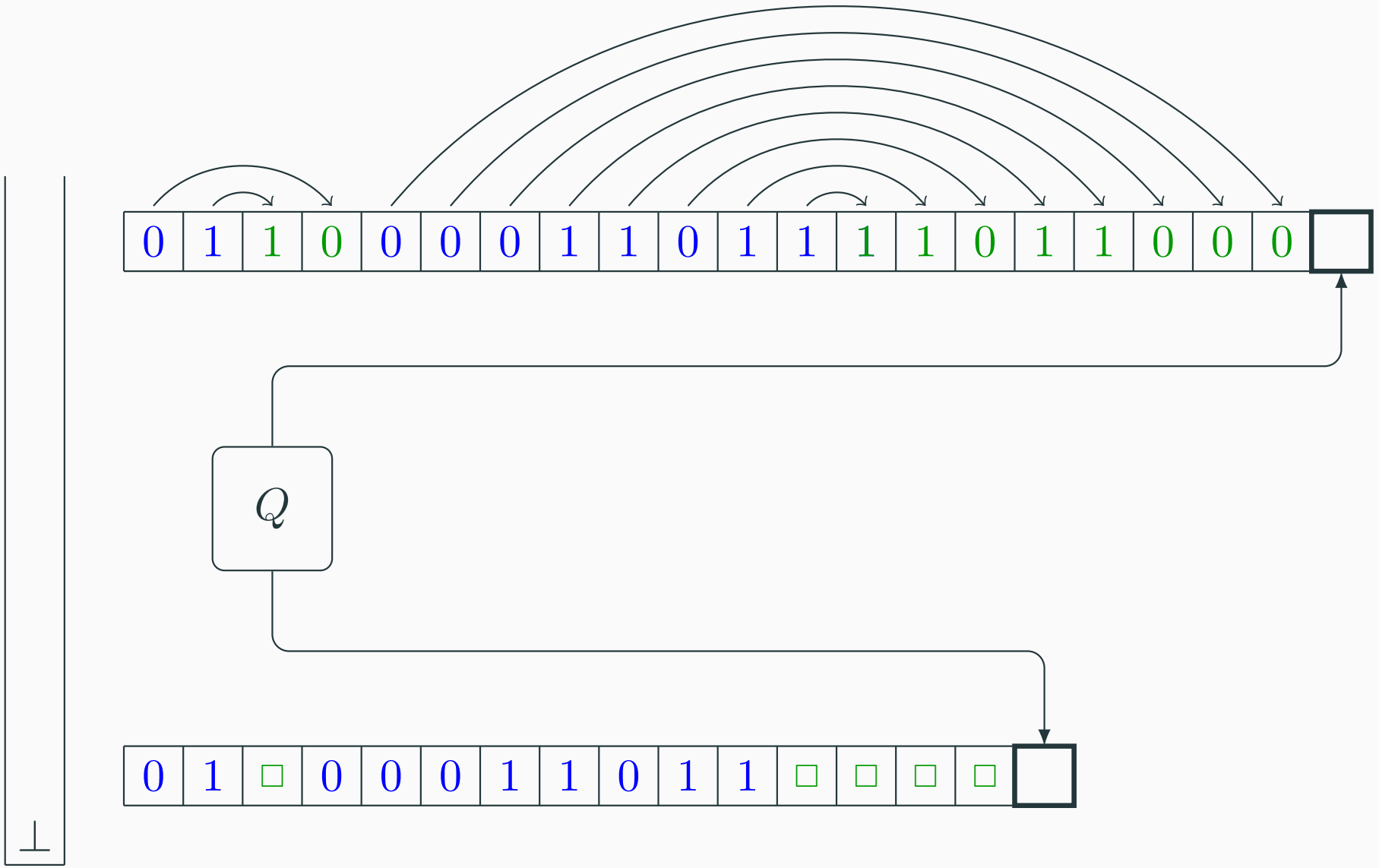
Let it run



Let it run



Let it run



Known result

Theorem

- *The transducer is one-to-one.*
- *The compression ratio is*

$$\frac{3\log(\#A+1)}{4\log(\#A)}.$$

Outline

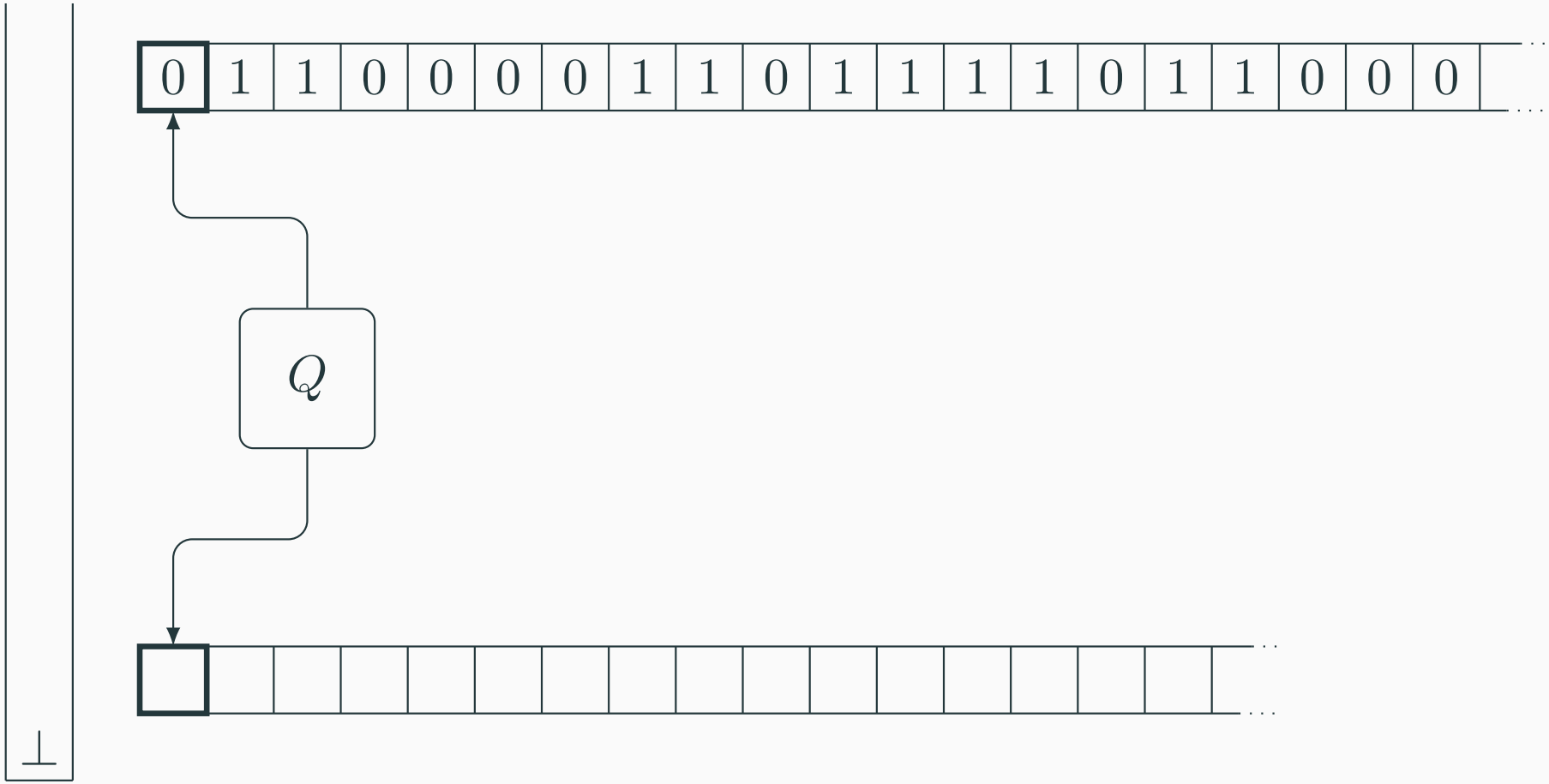
Deterministic pushdown transducers

$$x = w_1\tilde{w}_1 w_2\tilde{w}_2 w_3\tilde{w}_3 \cdots$$

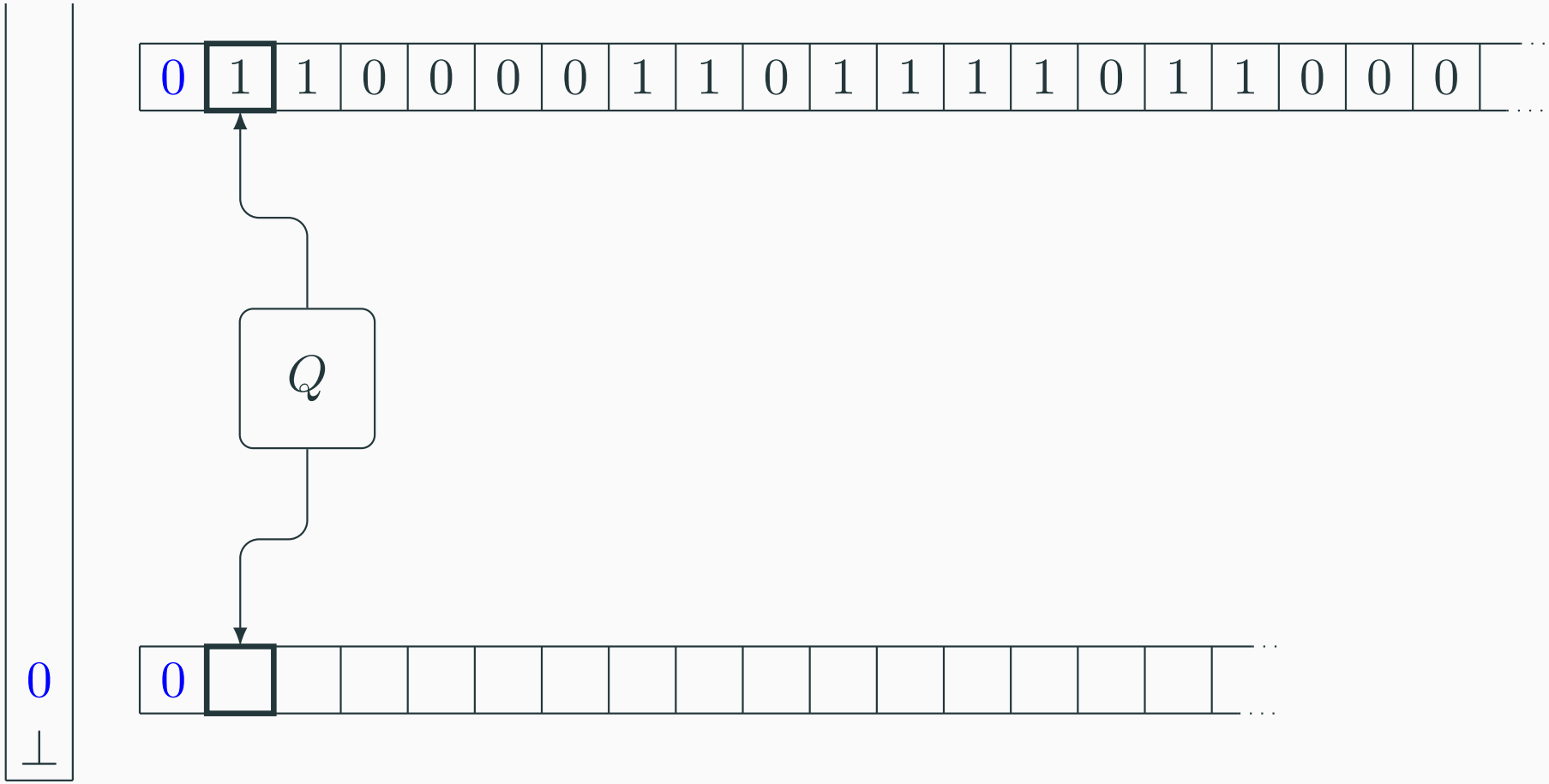
The transducer works as follows.

- If the input symbol is different from stack top symbol, the input symbol is pushed and output,
- If the input symbol and the stack top symbol are equal, the top symbol is popped.
  A symbol $\square$ is output every two popped symbols and a symbol $\triangle$ is output at the end if the number of popped symbols is odd.

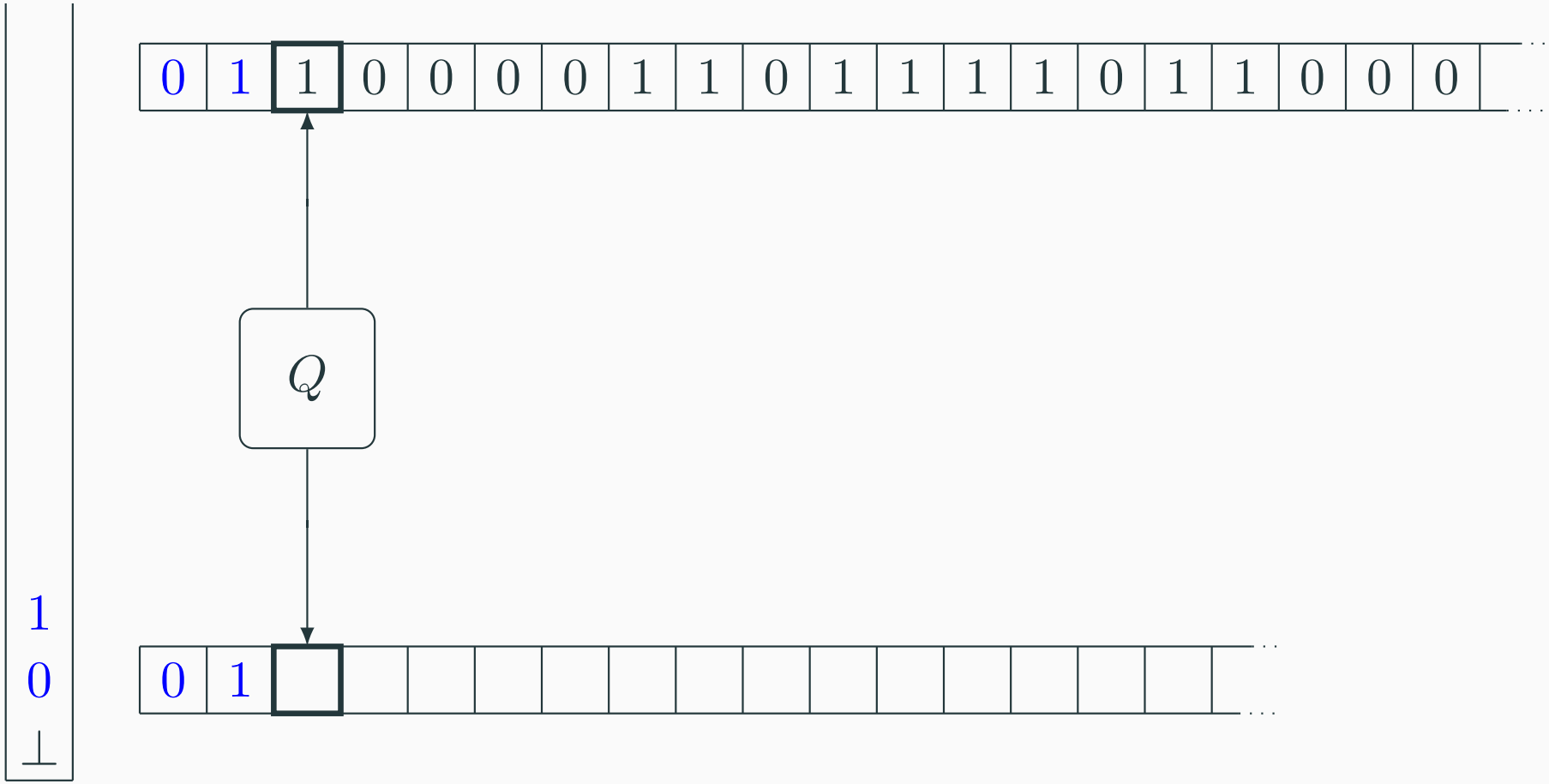
Let it run



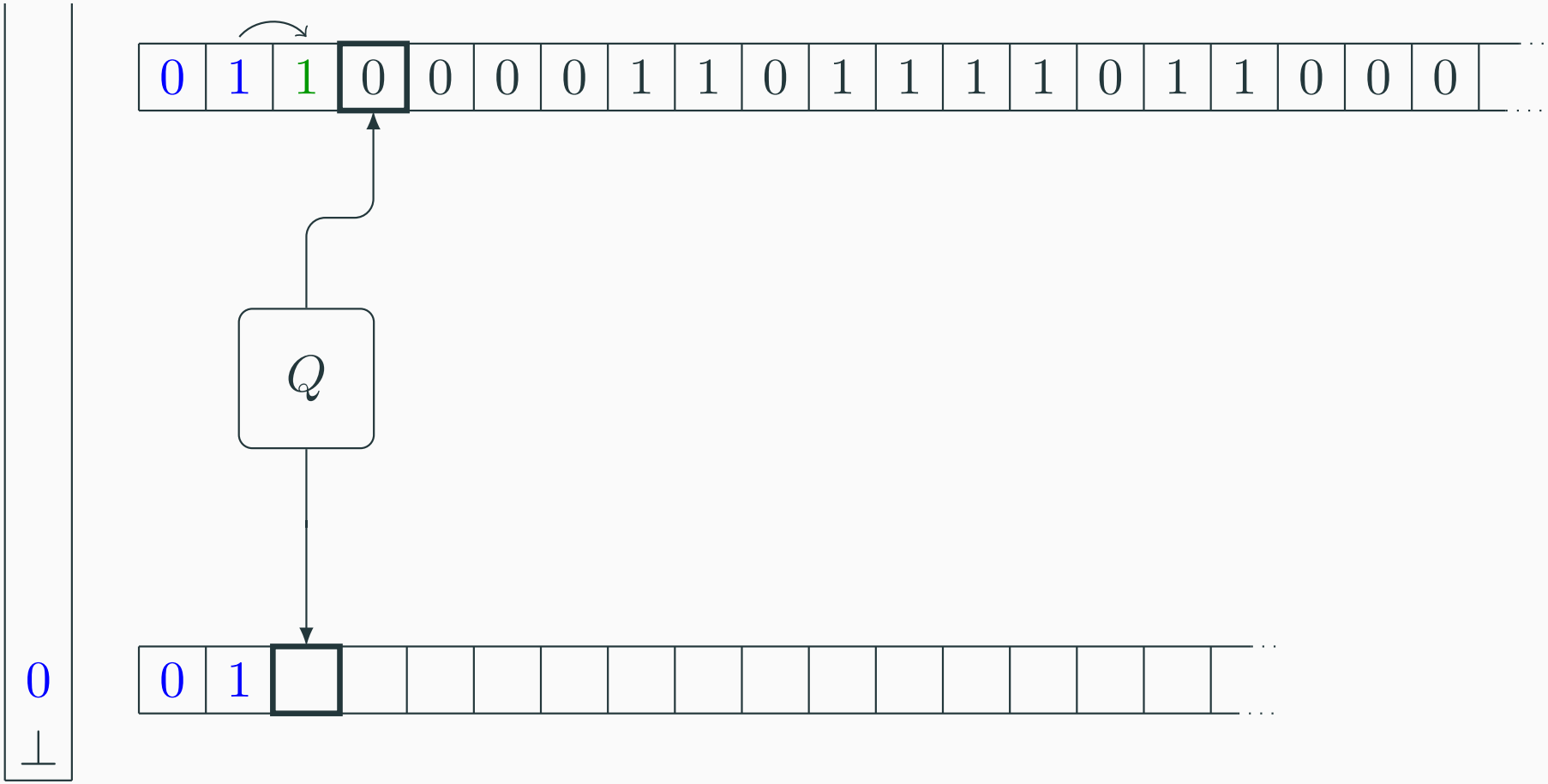
Let it run



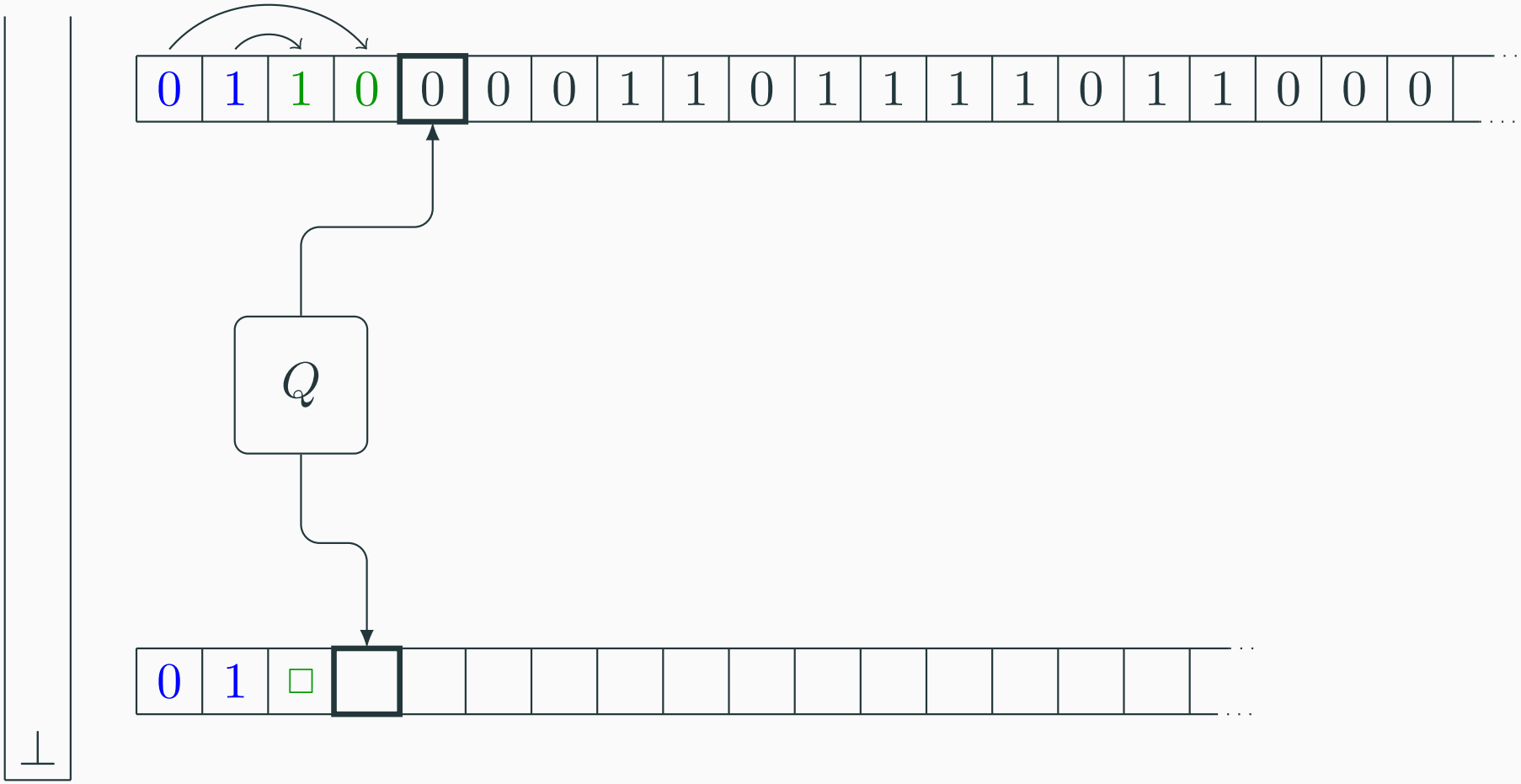
Let it run



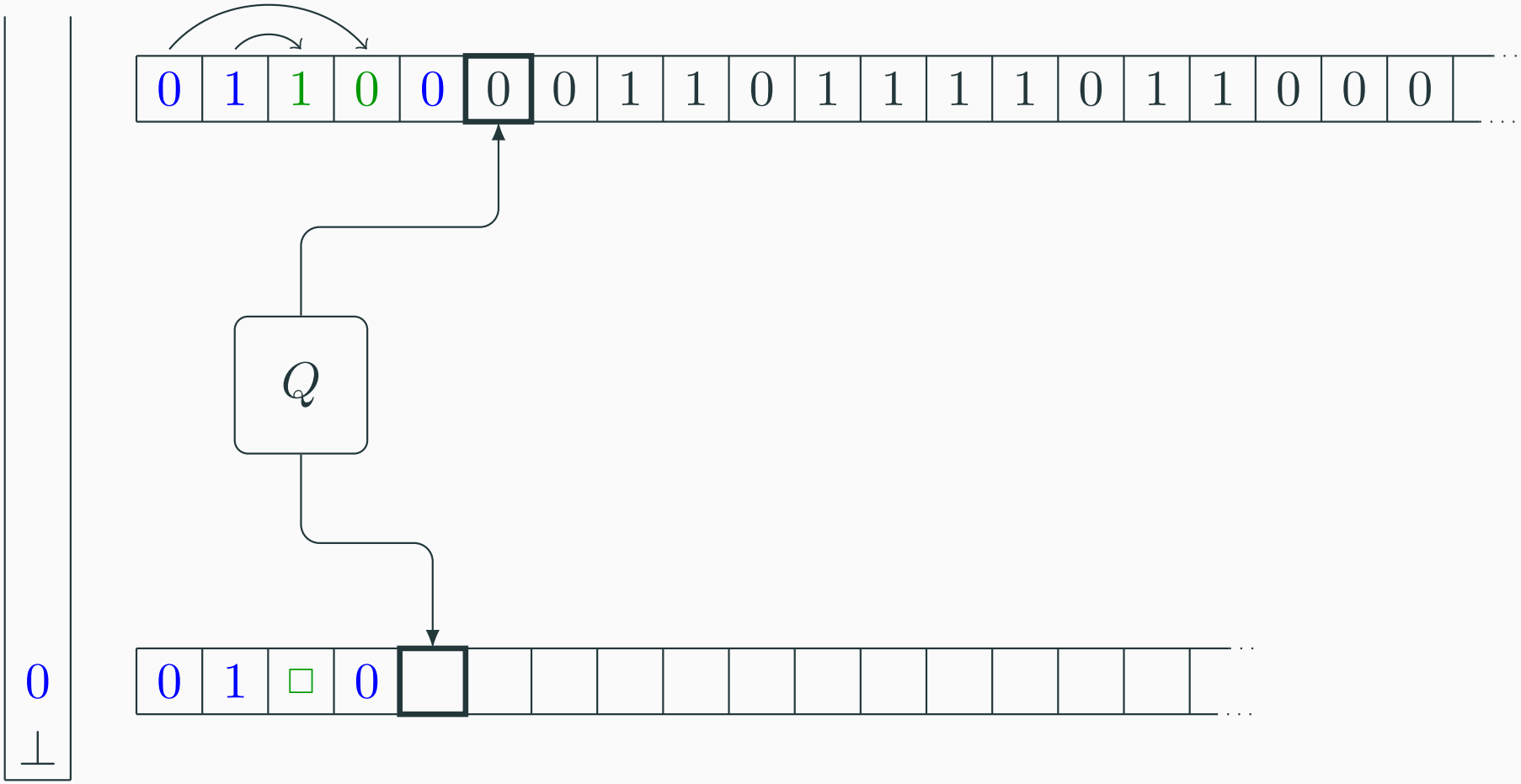
Let it run



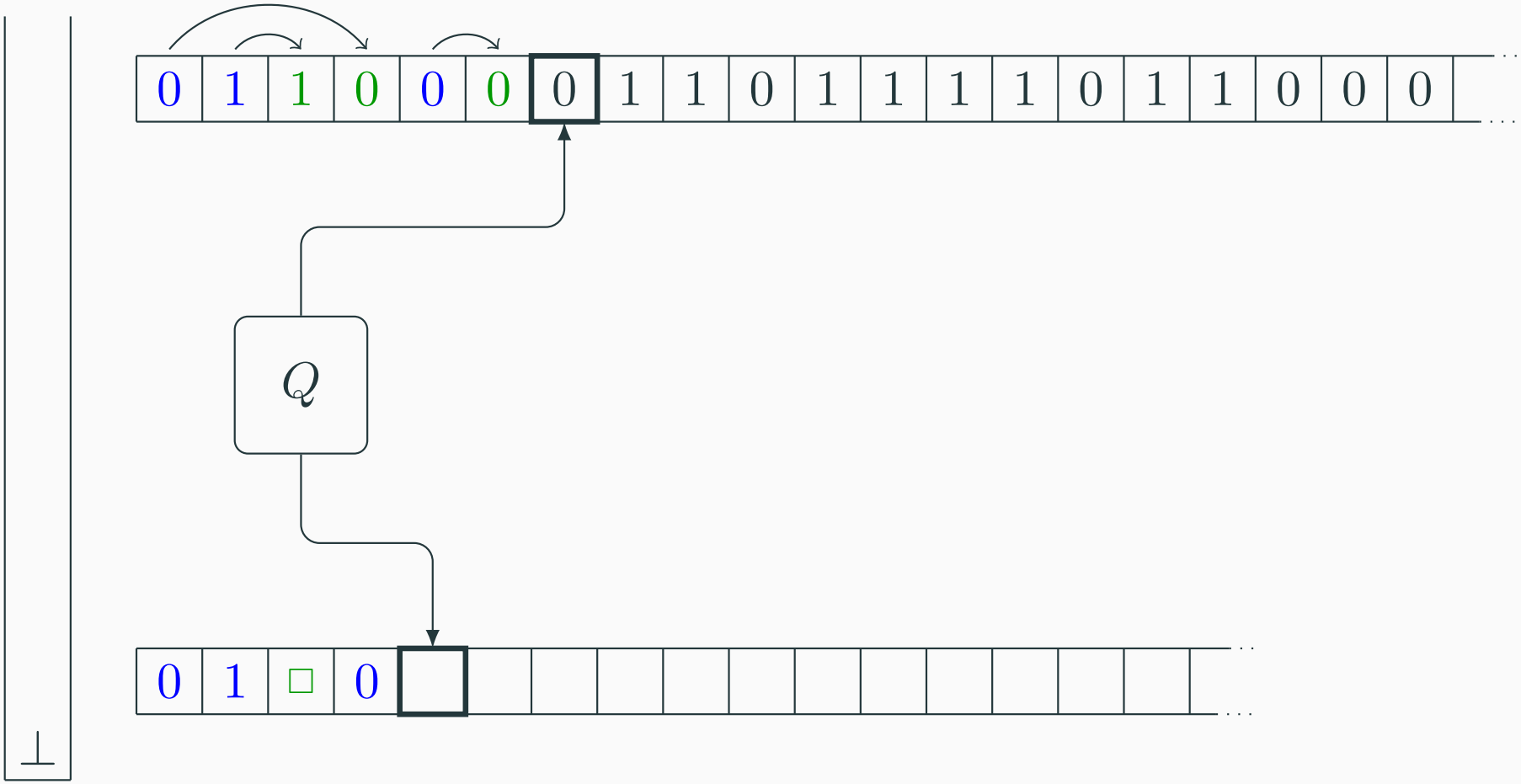
Let it run



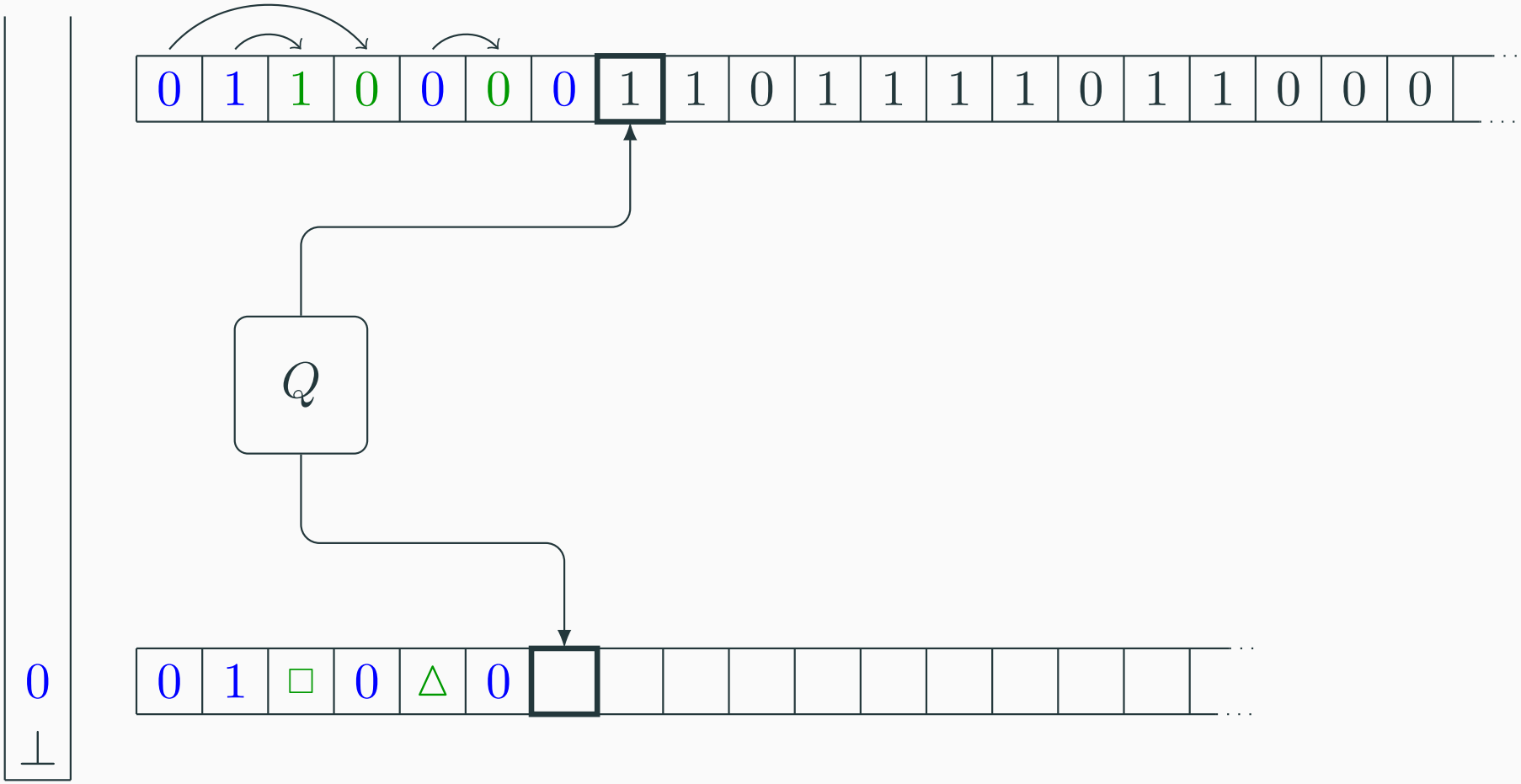
Let it run



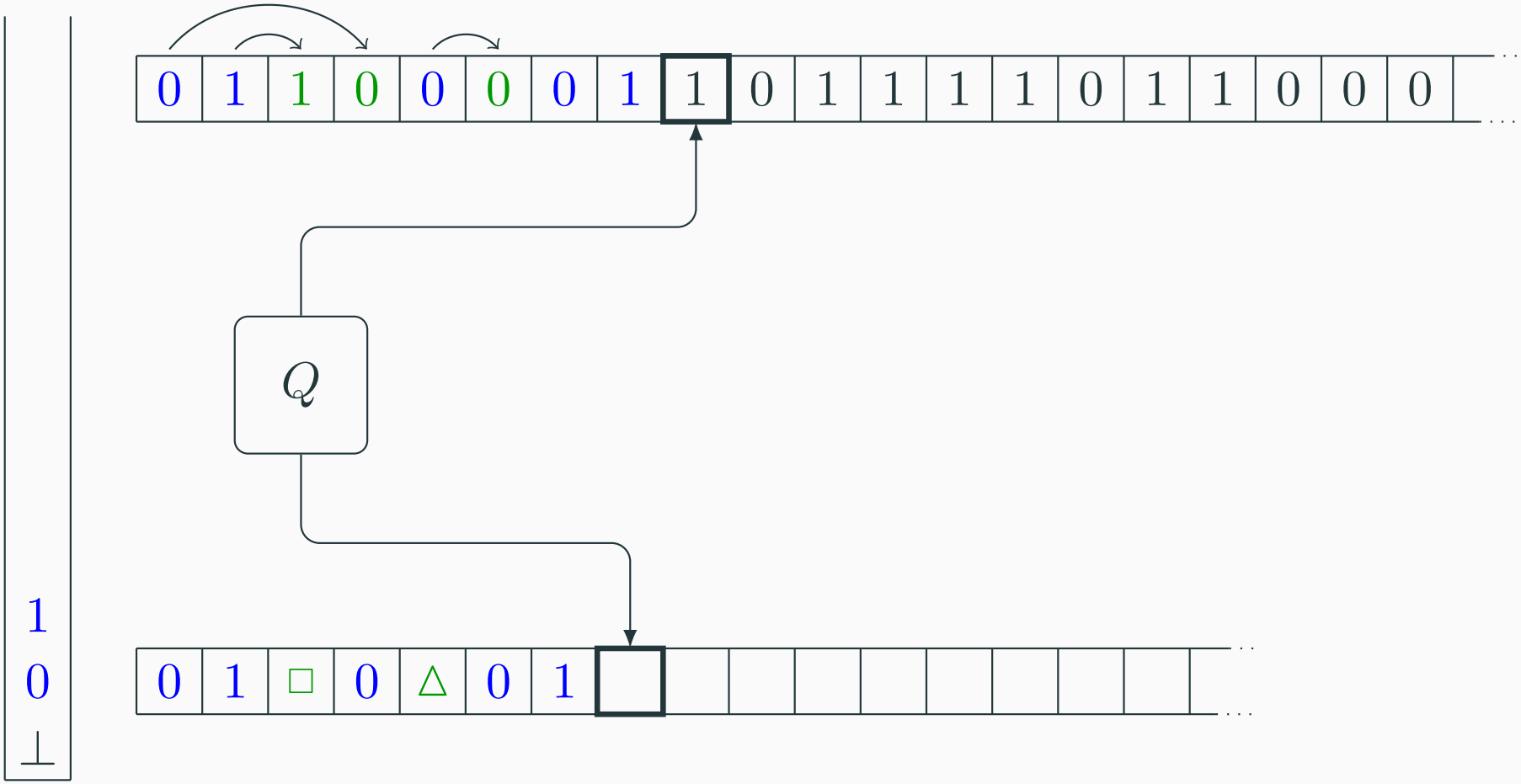
Let it run



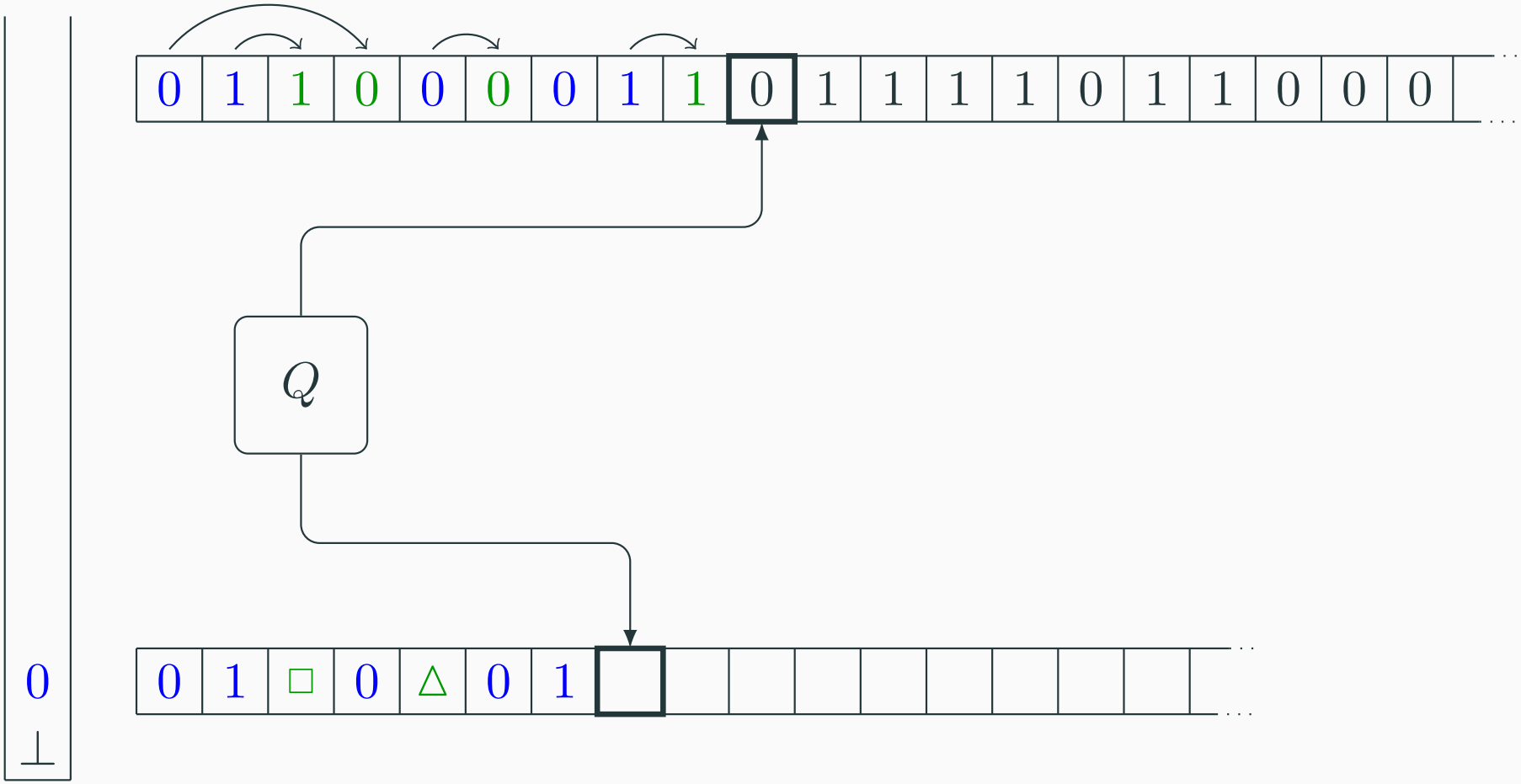
Let it run



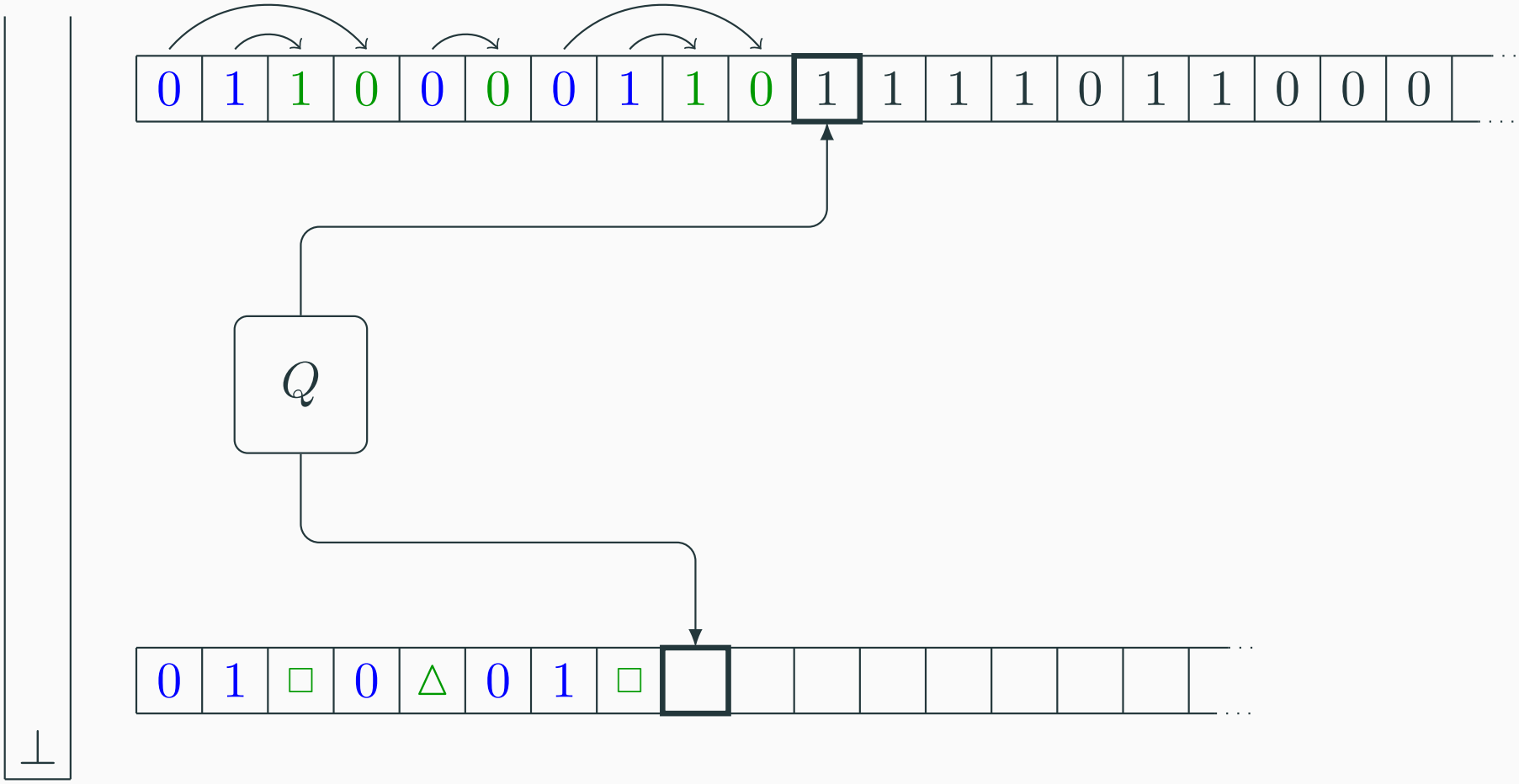
Let it run



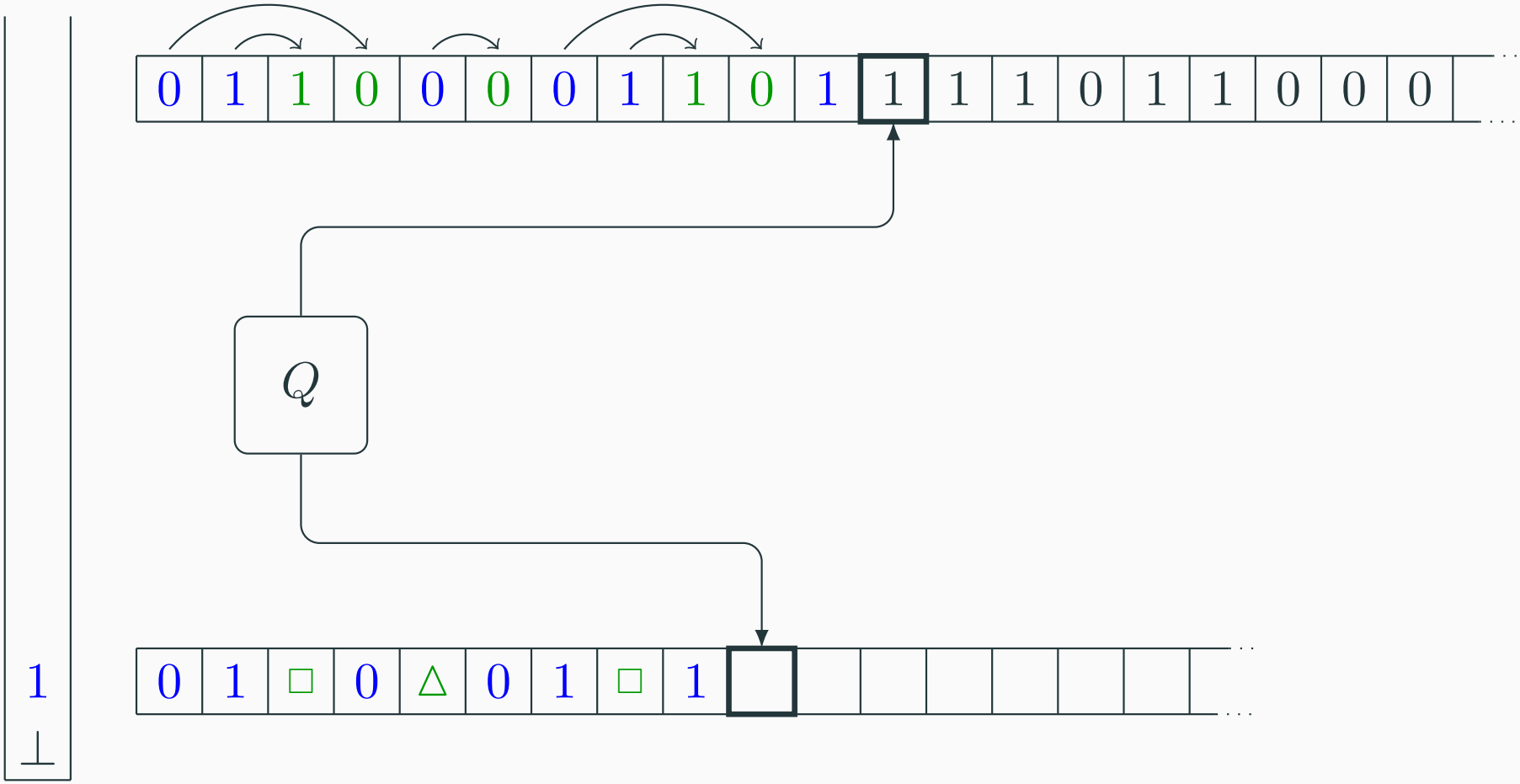
Let it run



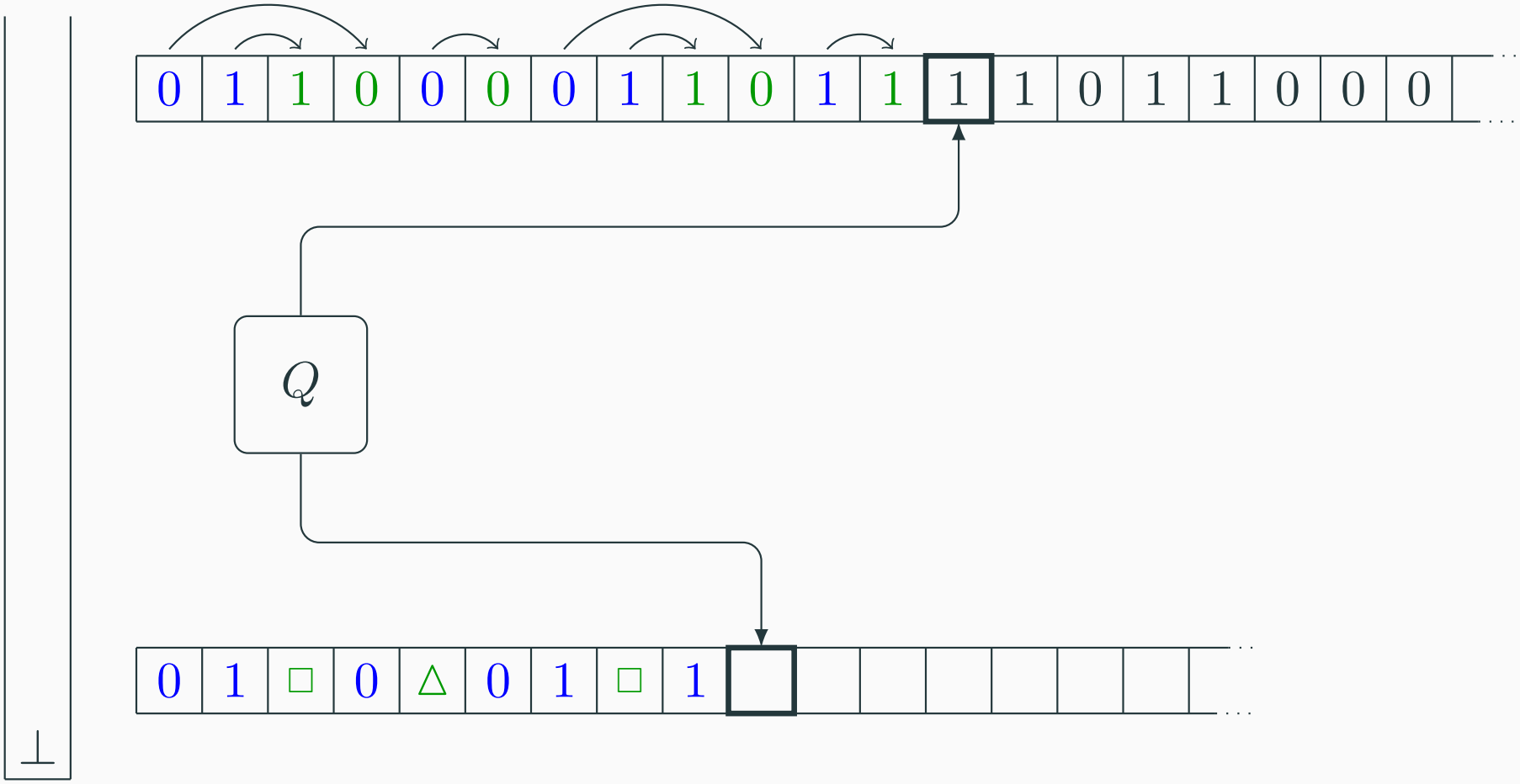
Let it run



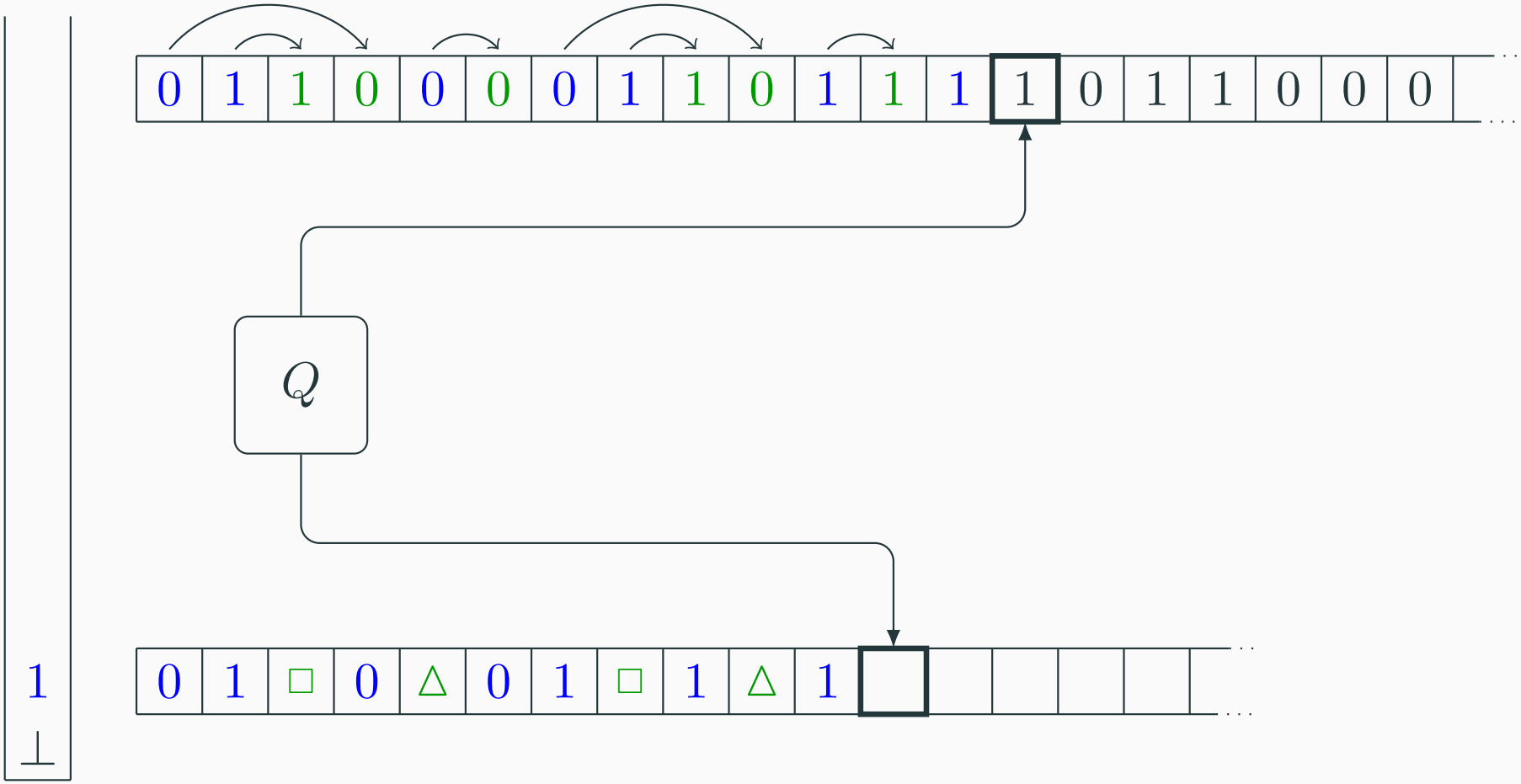
Let it run



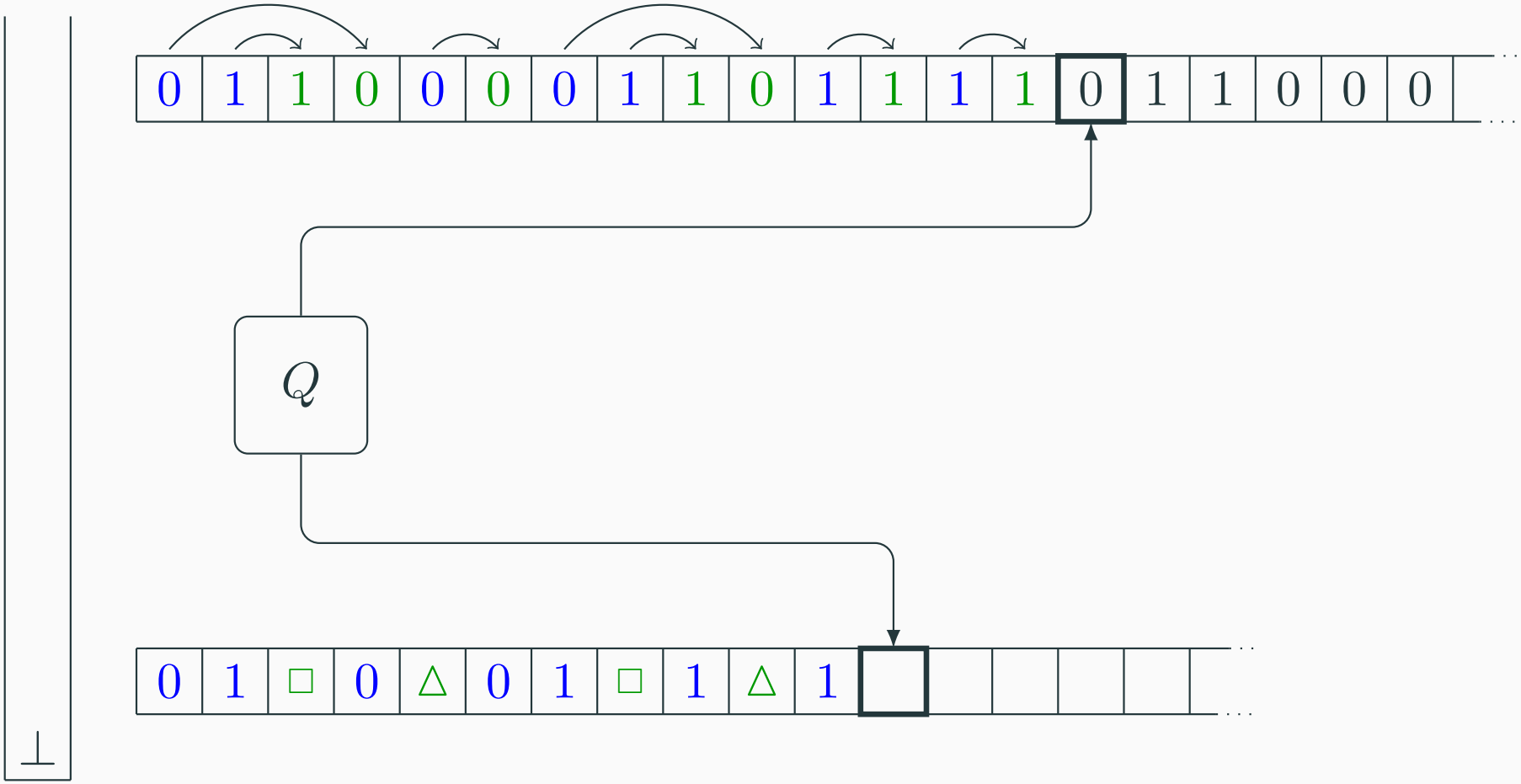
Let it run



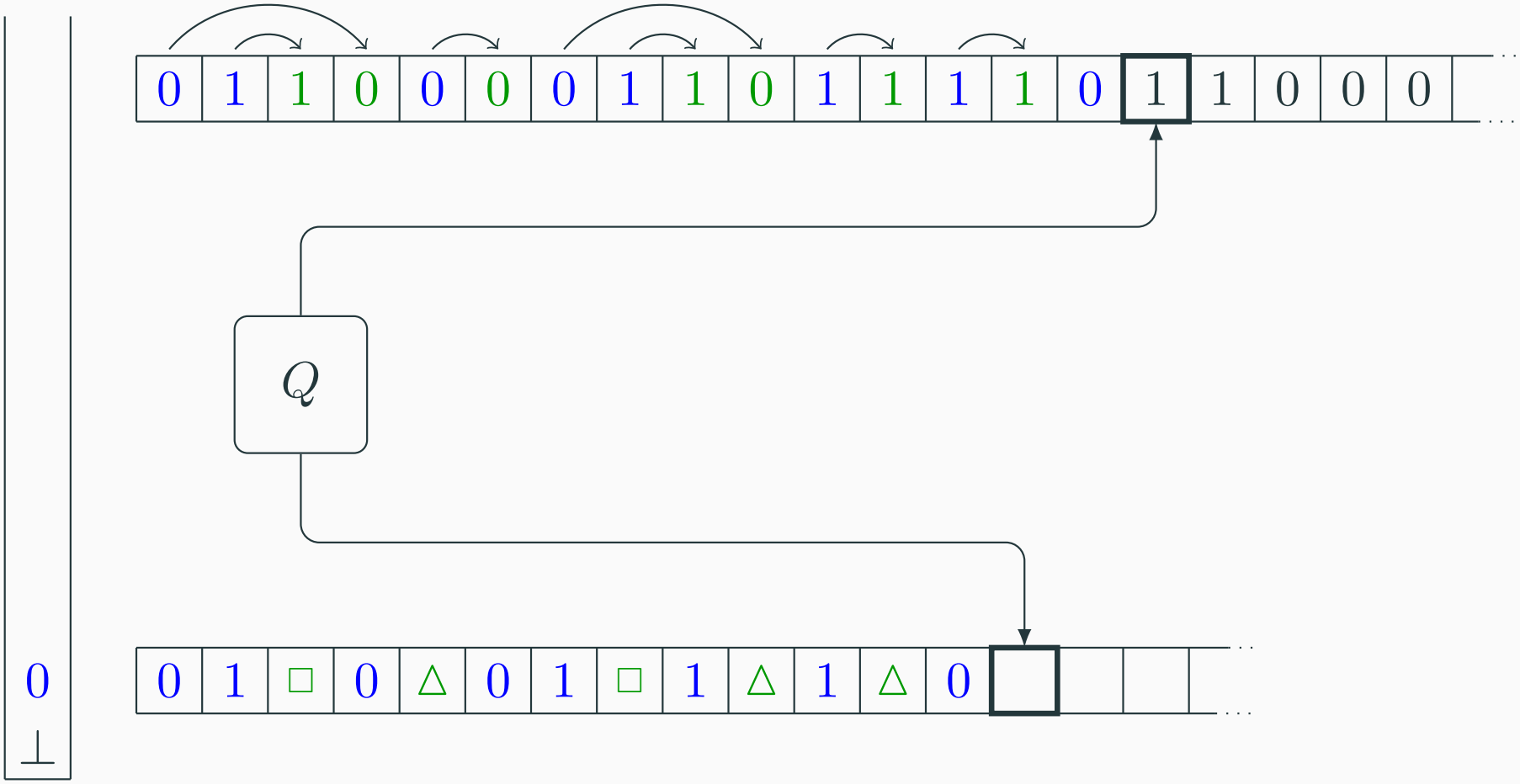
Let it run



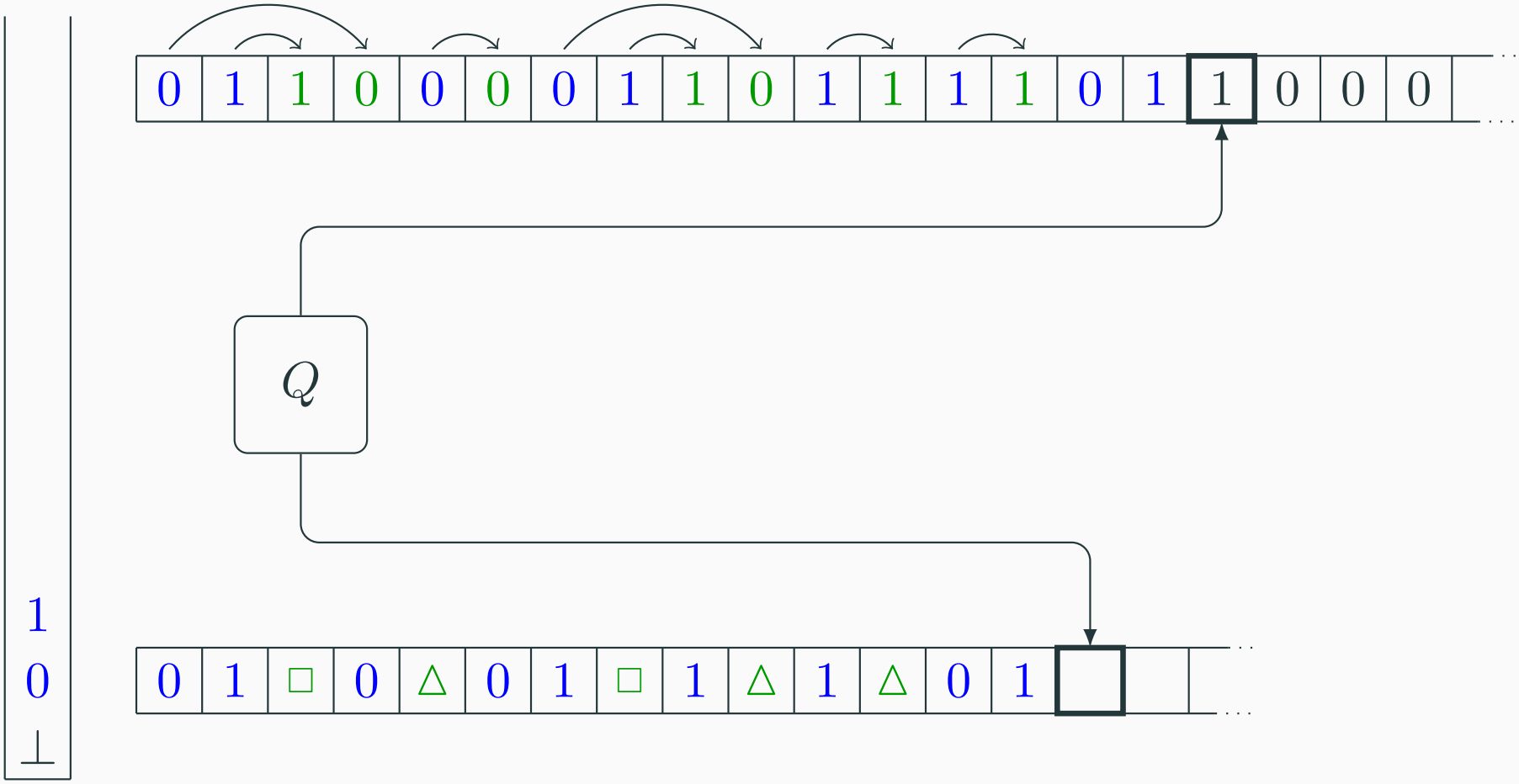
Let it run



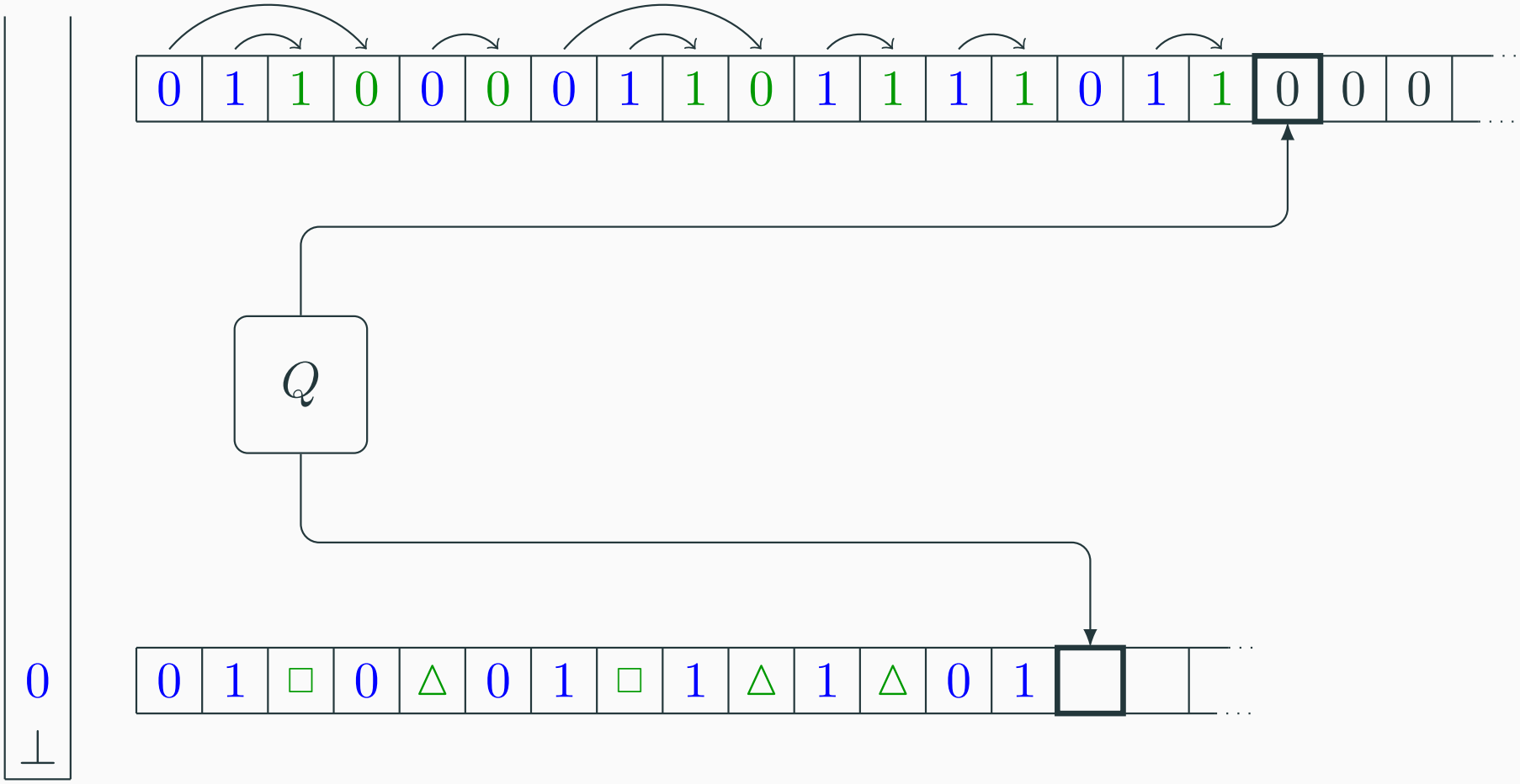
Let it run



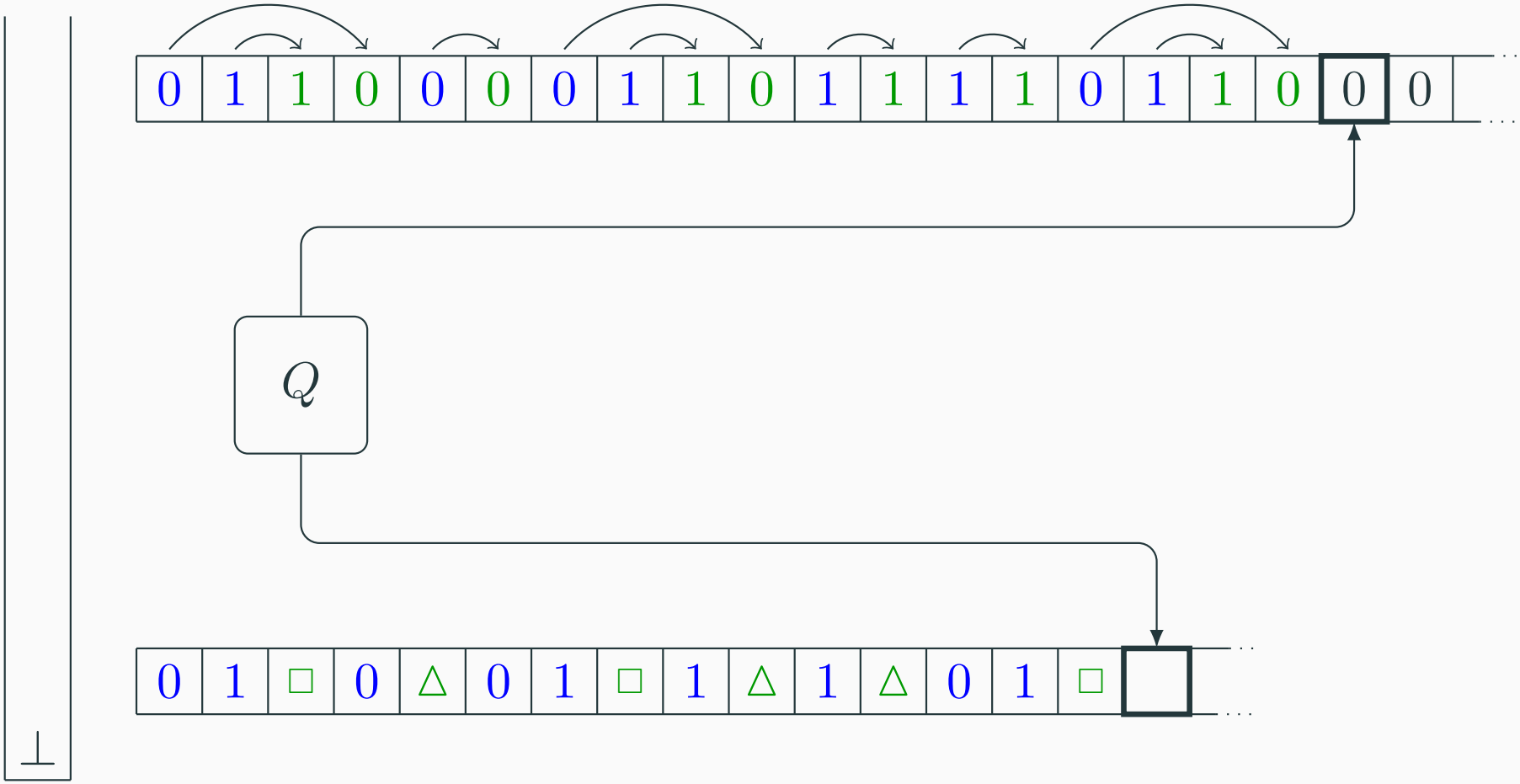
Let it run



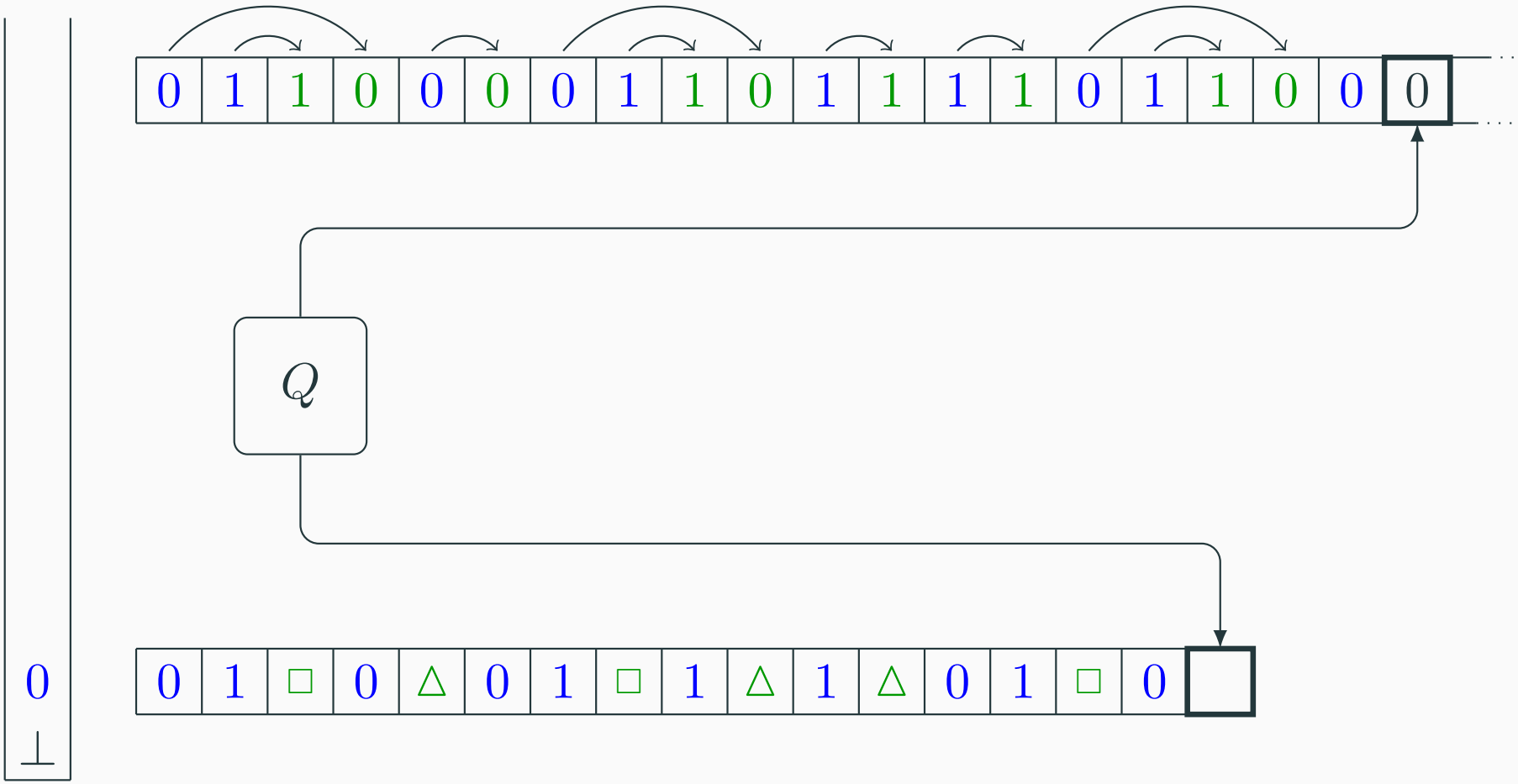
Let it run



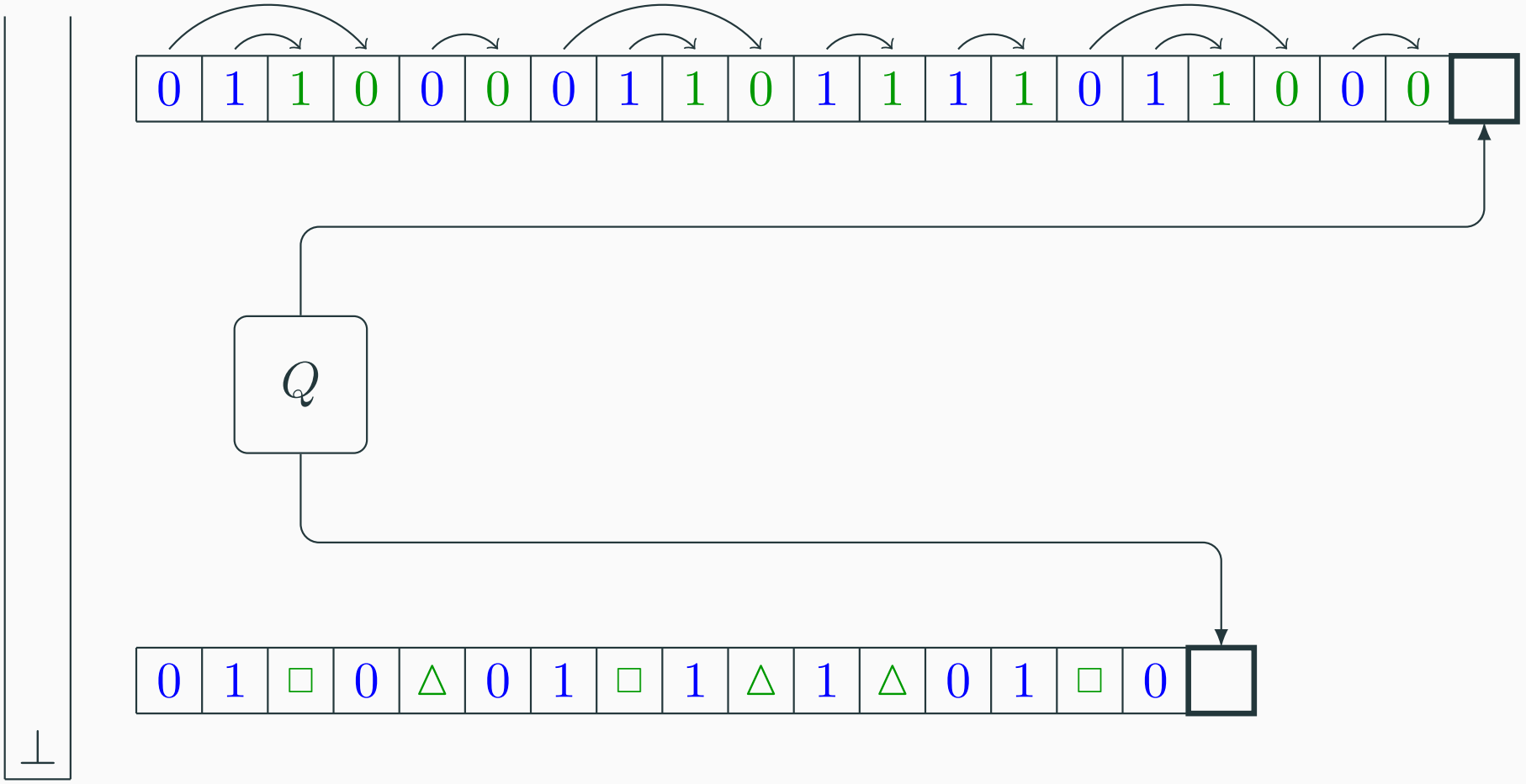
Let it run



Let it run



Let it run



Result

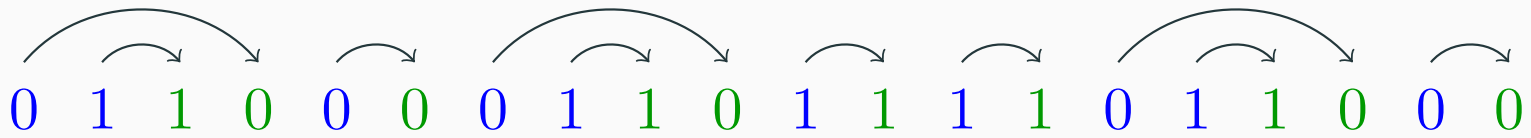
Theorem

- *The transducer is one-to-one.*
- *The compression ratio is less than*

$$\frac{19\log(\#A+2)}{20\log(\#A)}$$

for $\#A$ large enough.

Sketch of proof



- Each pair of consecutive pops saves one symbol.
- Each long edge implies a pair of consecutive pops.
- Each block of odd length implies a long edge.
- If $\#A$ is large, there are many blocks of length 1 in $w_n\tilde{w}_n$.

Open questions

- Which sequences are compressed by this transducer ?
- Does the compression ratio converge to $3/4$ when the alphabet size k goes to infinity ?
- ...

Merci