Reversible Computation

William Troiani

May 19, 2022

Computation, considered as an invisible mental process, is a naive stance. The crucial point is that there are genuine mathematical and physical properties of computation which are independent of the choice of implementation.

A historically significant indication that this might be the case, was the thought experiment often referred to as *Maxwell's Demon*, where it appears as though heat is being transfered from a cooler body to a warmer body, which contradicts the second law of thermodynamics. This thought experiment includes an onlooker (the demon) who stands idle and makes decisions. A proposed solution was that the decision process inside the demon's mind was to be included as part of the physical system. That is, the *computation* being performed by the Demon was a *physically* relevant part of the experiment.

Rigorous work following this thought experiment includes a result, first proved by Landauer, that irreversible computation is inherrently linked with physical irreversibility [1]. There, it is argued that practical computation fundamentally involves irreversible computation, however Bennett has proven this is not true [2]. In this note, we present Bennett's proof by showing that every terminating computation of a Turing machine can be simulated by a reversible Turing machine, see 0.0.5 for a precise statement.

What, though, makes a Turing machine M *reversible*? A crucial yet subtle point is that the resources required for M to produce output w' from input w may be wildly distinct from the resources required for some other Turing machine N to produce w from w'. For example, let M be a Turing machine which on input (M', y), consisting of a Turing machine M' and a valid output y for M', calculates an input z for M' so that when M' is run on z the output is y. This Turing machine M churns through all possible inputs, slowly increasing the time spent running on them, until such a z is eventually found. This is an extremely slow and expensive computation. However, to retrieve the input (M', y) from an output (M', z) we simply run M' on z.

Thus, we need a more refined notion of what it means for a Turing machine M to be *reversible*. In essence, we want there to exist another Turing machine N which performs every halting computation of M backwards. There will be many technical subtleties involved in creating such a definition, the first of which is a restriction on the head movements of a Turing machine.

Definition 0.0.1. A **Turing Machine** is a triple (Σ, Q, δ) consisting of:

- A finite alphabet (ie, a collection of symbols) Σ containing a special symbol '␣'.
- A finite set of **states** (again, symbols) Q containing special states $q_{\texttt{Start}}, q_{\texttt{Finish}}$.
- The **transition function**, ie, a function

$$\delta : \Sigma \times Q \longrightarrow \Sigma \times Q \times \{\texttt{Left}, \texttt{Stay}, \texttt{Right}\} \tag{1}$$

  The transition function is required to satisfy the following property: if $(\sigma, q) \in \Sigma \times Q$ is such that $\delta(\sigma, q) = (\sigma', q_{\texttt{Final}}, \texttt{Instr})$, for some $\sigma' \in \Sigma$, then $\texttt{Instr} = \texttt{Stay}$.

The set of finite words over the alphabet Σ will be denoted using the Kleene Star notation Σ^*.

Definition 0.0.2. A **computation** of a Turing Machine $M = (\Sigma, Q, \delta)$ on an input $w \in (\Sigma \setminus \{\sqcup\})^*$ is a (possibly infinite) sequence of configurations

$$(w_0, q_0, p_0) \longrightarrow (w_1, q_1, p_1) \longrightarrow \ldots \longrightarrow (w_i, q_i, p_i) \longrightarrow \ldots \tag{2}$$

satisfying the following conditions.

- **Initiality conditions:** The first triple (w_1, q_1, p_1) is subject to
  - $w_0 = w$.
  - $q_0 = q_{\texttt{Start}}$.
  - $p_0 = 0$. The words w_i are sequences starting at index 0, so this condition is saying that the Turing machine begins by reading the first letter in the input $w_1 = w$.
- **Movement conditions:** For all $i \geq 0$, denote by w_i^j the j^{th} letter in the word w_i. If $\delta(w_i^j, q_i) = (\overline{w_i^j}, \overline{q_i}, \texttt{Instr})$, then
  - $w_{i+1}^j = \overline{w_i^j}$.
  - $q_{i+1} = \overline{q_i}$.
  - and for p_i:

$$p_{i+1} = \begin{cases} p_i + 1 & \texttt{Instr} = \texttt{Right} \\ p_i & \texttt{Instr} = \texttt{Stay} \\ p_i - 1 & \texttt{Instr} = \texttt{Left} \end{cases} \tag{3}$$

- **Termination conditions:** There is at most one triple (w, q, p) such that $q = q_{\texttt{Finish}}$, and if there is one it is the final element. If this exists, we require also that $p = 0$.

The integer p_i is the **head position at step** i.

If the computation of a word w on a Turing Machine M is finite, then M **halts** on input w, we denote the output by $M(w)$.

Remark 0.0.3. Part of the termination conditions in Definition 0.0.2 are that the head position at step 1 and the final step (assuming the computation halts) is equal to 0. That is, the tape head position begins and ends at the start of the word.

This condition is not always given in the definition of a Turing machine, but it is necessary for reversible computation. The reason is as follows, if M is a Turing machine which halts on input w, then without this condition, the head position is some integer p. Thus, if we had a Turing machine N which "reversed" M, then N would need to start with head position at step 1 equal to p *which depends on the computation of* w *by* M.

To avoid this, we fix the initial and final head positions.

Definition 0.0.4. A Turing Machine $M = (\Sigma, Q, \delta)$ is **reversible** if there exists a transition function δ' (with domain equal to that of δ) and Turing machine $M^{-1} := (\Sigma, Q, \delta')$ subject to the following.

- If M halts on input w, ie, there is a finite computation

$$(w_0, q_0 = q_{\texttt{Start}}, p_0 = 0) \longrightarrow \ldots \longrightarrow (w_n, q_n = q_{\texttt{Finish}}, p_n = 0) \tag{4}$$

  then the computation of the Turing machine $M^{-1} := (\Sigma, Q, \delta')$ which begins at $(w_n, q_{\texttt{Finish}}, 0)$ terminates at $(w_0, q_{\texttt{Start}}, 0)$ and so there is a computation

$$(w_0' = w_n, q_0', p_0') \longrightarrow \ldots \longrightarrow (w_m', q_m', p_m') \tag{5}$$

  We require that this computation satisfies:

$$\text{For all } 0 \leq i \leq n, (w_i, q_i) = (w_{n-i}', q_{n-i}') \tag{6}$$

The amazing result, is the following.

Theorem 0.0.5. *For every Turing machine $M = (\Sigma, Q, \delta)$ there exists a reversible Turing machine $N = (\Sigma', Q', \delta')$ satisfying the following.*

- $\Sigma \subseteq \Sigma$.
- *If M terminates on input w with output $M(w)$ then N terminates on input w with output $w\#\#\#M(w)$, where $\#$ is some distinguished letter in Σ'.*

The proof is inspired by the standard way of making a non-injective function injective: given arbitrary $f : X \longrightarrow Y$ we define $\hat{f} : X \longrightarrow X \times Y$ which sends $x \longrightarrow (x, f(x))$. That is, $\hat{f}$ is the *graph* of f.

Similarly, we will add two tapes to M which will be used to write down a "history" of the steps of δ performed. The machine N then uses this "history" tape to unwind the performance of M, after adding a fourth tape to write down the output of M.

Lemma 0.0.6. *To every Turing machine $M = (\Sigma, Q, \delta)$ there is a 4-tape reversible Turing machine N so that for every input w upon which M halts, the machine N also halts, and the final position of N has w written on the first tape, the second and third tape is blank, and has $M(w)$ written on the fourth tape.*

Example 0.0.7. Say we had a Turing machine M which performed the following simple computation, where an underlined symbol denotes the position of the head.

$$\begin{array}{ccccccccc} & q_{\texttt{Start}} & & q_1 & & q_2 & & q_3 & & q_{\texttt{Finish}} \\ |\underline{\sigma_1}|\sigma_2|\ldots & \longrightarrow & |\overline{\sigma_1}|\underline{\sigma_2}|\ldots & \longrightarrow & |\overline{\sigma_1}|\underline{\overline{\sigma_2}}|\ldots & \longrightarrow & |\underline{\overline{\sigma_1}}|\overline{\overline{\sigma_2}}|\ldots & \longrightarrow & |\underline{\overline{\overline{\sigma_1}}}|\overline{\overline{\sigma_2}}|\ldots \end{array} \tag{7}$$

This has its history recorded as follows.

$$\begin{array}{ccccccccc} & q_{\texttt{Start}} & & q_1 & & q_2 & & q_3 & & q_{\texttt{Finish}} \\ |\underline{\sigma_1}|\sigma_2|\ldots & \longrightarrow & \overline{\sigma_1}|\underline{\sigma_2}|\ldots & \longrightarrow & |\overline{\sigma_1}|\underline{\overline{\sigma_2}}|\ldots & \longrightarrow & |\underline{\overline{\sigma_1}}|\overline{\overline{\sigma_2}}|\ldots & \longrightarrow & |\underline{\overline{\overline{\sigma_1}}}|\overline{\overline{\sigma_2}}|\ldots \\ & & \big((\sigma_1, q_{\texttt{Start}}), (\overline{\sigma_1}, q_1)\big) & & \big((\sigma_2, q_1), (\overline{\sigma_2}, q_2)\big) & & \big((\overline{\sigma_2}, q_2), (\overline{\overline{\sigma_2}}, q_3)\big) & & \big((\overline{\sigma_1}, q_3), (\overline{\overline{\sigma_1}}, q_{\texttt{Finish}})\big) \\ & & \texttt{Right} & & \texttt{Stay} & & \texttt{Left} & & \texttt{Stay} \end{array}$$

We then delete the final instance of `Stay` and move the contents of the third tape one square to the right, adding an instance of `Stay` to the beginning.

$$\begin{array}{ccccccccc} & q_{\texttt{Start}} & & q_1 & & q_2 & & q_3 & & q_{\texttt{Finish}} \\ |\underline{\sigma_1}|\sigma_2|\ldots & \longrightarrow & |\overline{\sigma_1}|\underline{\sigma_2}|\ldots & \longrightarrow & |\overline{\sigma_1}|\underline{\overline{\sigma_2}}|\ldots & \longrightarrow & |\underline{\overline{\sigma_1}}|\overline{\overline{\sigma_2}}|\ldots & \longrightarrow & |\underline{\overline{\overline{\sigma_1}}}|\overline{\overline{\sigma_2}}|\ldots \\ & & \big((\sigma_1, q_{\texttt{Start}}), (\overline{\sigma_1}, q_1)\big) & & \big((\sigma_2, q_1), (\overline{\sigma_2}, q_2)\big) & & \big((\overline{\sigma_2}, q_2), (\overline{\overline{\sigma_2}}, q_3)\big) & & \big((\overline{\sigma_1}, q_3), (\overline{\overline{\sigma_1}}, q_{\texttt{Finish}})\big) \\ & & \texttt{Stay} & & \texttt{Right} & & \texttt{Stay} & & \texttt{Left} \end{array}$$

We then "turn these instructions around", but doing the opposite to what each step says.

$$\begin{array}{cccc} \big((\overline{\overline{\sigma_1}}, q_{\texttt{Finish}}), (\overline{\sigma_1}, q_3)\big) & \big((\overline{\overline{\sigma_2}}, q_3), (\overline{\sigma_2}, q_2)\big) & \big((\overline{\sigma_2}, q_2), (\sigma_2, q_1)\big) & \big((\overline{\sigma_1}, q_1), (\sigma_1, q_{\texttt{Start}})\big) \\ \texttt{Right} & \texttt{Stay} & \texttt{Left} & \texttt{Stay} \end{array} \tag{8}$$

This, when acted on input $|\overline{\overline{\sigma_1}}|\overline{\overline{\sigma_2}}|\dots$ yields the following computation.

$$\begin{array}{ccccccccc} & q_{\texttt{Finish}} & & q_3 & & q_2 & & q_1 & & q_{\texttt{Start}} \\ |\overline{\overline{\sigma_1}}|\overline{\overline{\sigma_2}}|\dots & \longrightarrow & |\overline{\sigma_1}|\overline{\overline{\sigma_2}}|\dots & \longrightarrow & |\overline{\sigma_1}|\overline{\sigma_2}|\dots & \longrightarrow & |\overline{\sigma_1}|\sigma_2|\dots & \longrightarrow & |\underline{\sigma_1}|\sigma_2|\dots \end{array} \tag{9}$$

This is exactly the reverse of (7), up to head position.

Lemma 0.0.8. *Let M be an n-tape Turing machine $M = (\Sigma, Q, \delta)$. There exists a (1-tape) Turing machine N satisfying the following property.*

- *For every input $(w_1, \dots, w_n)$ of M for which M terminates, say with output $(o_1, \dots, o_n)$ we have $N(w_1\# \dots \# w_n) = o_1\# \dots \# o_n$, where $\#$ is some distinguished letter in the alphabet of N.*

Proof. We will compress all the tapes onto a single square and initiate a "read" phase followed by a "write" phase.

Define the following alphabet

$$\Sigma' := (\Sigma \times \{\texttt{Y}, \texttt{N}\})^n \tag{10}$$

The $\texttt{Y}$ is used to denote the position of the head for this particular tape.

The states need to pick up the slack of only having a single tape and so is more complicated than the alphabet.

$$Q' := \Big((\Sigma \coprod \{?\})^n \times (Q \coprod \{q_{\texttt{Read}}\}\Big) \coprod \Big((\Sigma \coprod \{?\})^n \times Q \times \{\texttt{Left}, \texttt{Stay}, \texttt{Right}, \texttt{Update}\}^n\Big) \tag{11}$$

The starting state is $((?, \dots, ?), q_{\texttt{Start}})$ and the final state is $((?, \dots, ?), q_{\texttt{Finish}})$.

The transition function is defined so that the first symbol on the first square is read and then the state transitions into a state with second component $q_{\texttt{Read}}$. If a symbol of the form

$$\big((\sigma_1, \texttt{N}), \dots, (\sigma_i, \texttt{Y}), \dots, (\sigma_n, \texttt{N})\big) \tag{12}$$

is read, the i^{th} question mark symbol '?' in the first component of the state is changed to σ_i.

Once the first component of the state has no more question mark symbols, we transition into the "update" phase.

The state is of the form $\big((\sigma_1, \dots, \sigma_n), q_{\texttt{Read}}\big)$. We calculate $\delta((\sigma_1, \dots, \sigma_n), q_{\texttt{Start}})$ and store the solution in the state. That is, if

$$\delta((\sigma_1, \dots, \sigma_n), q_{\texttt{Start}}) = \big((\overline{\sigma_1}, \dots, \overline{\sigma_n}), q, (\texttt{Instr}_1, \dots, \texttt{Instr}_n)\big) \tag{13}$$

then we update the state to exactly this tuple.

The Turing machine N then passes from the right of the tape to the left updating the symbols. That is, if we come across a symbol of the form (12), the machine N rewrites the portion of the symbol $(\sigma_i, \texttt{Y})$ as $(\overline{\sigma_i}, \texttt{N})$ if the i^{th} element of the third entry of (13) is $\texttt{Right}$ or $\texttt{Left}$. The $\texttt{Y}$ is left unchanged otherwise.

Then the neighbouring tuple is updated to mark a $\texttt{Y}$ where the head on the n-tape Turing machine moved to. The $\texttt{Instr}_i$ components of the state are changed to $\texttt{Update}$ while this happens. If there are two portions of the symbol in (12) with a $\texttt{Y}$, then we update the one which updates to the *right* first. If this is none of them then we move to the left and update all of them.

Once the head is at the far left of the tape contents, the state is updated to $((?, \dots, ?), q_{\texttt{Start}})$ and the "reading" phase of the Turing machine is reinitiated.

We leave it to the reader to prove that from this, the output can be converted into the form appropriate for the Lemma. □

Lemma 0.0.9. *If an n-tape Turing machine is reversible, then so is the simulating (1-tape) Turing machine.*

Proof. All the intermediate steps of reading and rewriting are clearly reversible, the only non-trivial part is when the simulating machine computes $\delta((\sigma_1, \ldots, \sigma_n), q) = ((\overline{\sigma_1}, \ldots, \overline{\sigma_n}), \overline{q}, (\texttt{Instr}_1, \ldots, \texttt{Instr}_n))$, we laboured the point that a reversible Turing machine is not equivalent to asking for an injective transition function, so we need to know how to restore this input from this output. The part of the transition function is applied *in a computation*, so the existence of a reverse means there is another transition function which although may not be an inverse to δ, at least calculates $((\sigma_1, \ldots, \sigma_n), q)$ from $((\overline{\sigma_1}, \ldots, \overline{\sigma_n}), \overline{q})$. Thus we can construct a reverse to N (the simulating Turing machine). □

References

[1] R. Landauer, *Irreversibility and heat generation in the computing process*

[2] C. H. Bennet, *Logical reversibility of computation*

[3] J. Clift, *Universal Turing Machines*, `http://therisingsea.org/notes/talk-james-utm.pdf`

[4] W. Troiani, *Reversible Turing Machines*, `http://therisingsea.org/notes/talk-will-reversible.pdf`