A conversation with Professor David Wolpert, January 24, 2020

Participants

- Professor David Wolpert -- Resident Faculty at the Santa Fe Institute
- Joseph Carlsmith -- Research Analyst, Open Philanthropy

Note: These notes were compiled by Open Philanthropy and give an overview of the major points made by Prof. Wolpert.

Summary

Open Philanthropy spoke with Prof. David Wolpert as part of its investigation of what we can learn from the brain about the computational power ("compute") sufficient to match human-level task performance. The conversation focused on the applicability of Landauer's principle to the brain's computation.

Landauer's Principle

Landauer's principle, as originally formulated by Rolf Landauer in 1961, states that erasing a single bit of information requires a minimum energy expenditure -- specifically, kT ln 2, where k is Boltzmann's constant and T is the absolute temperature. Landauer's original formula applies to a simple case in which there are two possible states, a uniform probability distribution over the states, and one heat bath. These are special conditions, and if they don't hold, you need different formulas.

A lot of progress in understanding the physics of computation has been made since 1961. At that point, physicists like Landauer only had the tools of equilibrium statistical physics available. But they were applying these tools to highly non-equilibrium systems, like computers. Consequently, their results, while based on sound intuition, were informal, imprecise, slightly mistaken, and limited to very simple computations like erasing bits.

Physicists have since made breakthroughs in nonequilibrium statistical physics (sometimes called stochastic thermodynamics). In particular, they have formulated the generalized Landauer bound, which applies to a broader set of conditions, and which reflects a much more detailed understanding of the relevant theory.

Logical and thermodynamic reversibility

There is a huge amount of confusion, in popular discourse about Landauer's principle, about the relationship between logical and thermodynamic reversibility. These are two completely distinct concepts. Logical reversibility concerns the dynamics of a single, particular state. Thermodynamic reversibility concerns the dynamics of a distribution over states. You can perform logically irreversible operations, like bit-erasures, in a thermodynamically reversible way, and you can perform logically reversible operations in a thermodynamically irreversible way.

The generalized Landauer bound tells you the energy costs of performing a computation in a thermodynamically reversible way -- energy that you could in principle get back. In particular: if you're connected to a single heat bath, then regardless of whether your computation is deterministic or noisy, the generalized Landauer's bound says that the minimum free energy you need to expend (assuming you perform the computation in a thermodynamically reversible way) is kT multiplied by the drop in the entropy.

The *total* energy costs of a computation will then be the Landauer cost, *plus* the extra energy dissipated via the thermodynamically irreversible aspects of the physical process. This extra energy cannot be recovered.

For a clear and relatively accessible exposition of this distinction, Prof. Wolpert suggests the paper "Thermodynamic and Logical Reversibilities Revisited" by Prof. Takahiro Sagawa of the University of Tokyo.

Reversible computing

Prof. Wolpert does not think that using logically reversible operations allows you to bypass the Landauer bound, assuming you want to re-use the reversible computer in question. Reversible computing requires storing enough information to undo the computation. If you want to re-use the computer, however, you need to return it to its initial state, which means you have to erase everything you store along the way, and you have to erase the initial input in order to get a new one. This wipes out any energy savings.

Much of the original work on reversible computation was done prior to recent advances in non-equilibrium statistical physics, and it did not incorporate the formal rigor that new tools make available. Contemporary statistical physicists do not really pay attention to logically reversible computation, because they view it as a non-starter, and top-tier physics journals rarely publish papers on the topic.

Thermodynamics and neurobiology

Metabolic constraints are extremely important in evolutionary biology. But the field of evolutionary biology has not adequately incorporated discoveries about the energy costs of the computation.

The massive energy costs of the brain ground a presumption that it has been highly optimized for thermodynamic efficiencies. Understanding better how the brain's architecture balances energy costs with computational performance may lead to important breakthroughs. However, at this point we are basically clueless about how the brain's computation works, so we can't even state this problem precisely.

Applicability of Landauer's principle to the brain

Mr. Carlsmith asked Prof. Wolpert whether one can use Landauer's principle to upper bound the FLOP/s required to replicate the human brain's task-performance. The argument in question would proceed by first estimating the number of bit-erasures the brain could be performing per second, by dividing its energy consumption (~20W) by kT ln 2. This calculation outputs around 1e22 bit-erasures per second, according to Mr. Carlsmith's estimates. The next step would be to translate this into a cap on the FLOP/s required to replicate to the brain's task-performance overall.

In Prof. Wolpert's view, it is a subtle and interesting question how to do this type of calculation correctly. A rigorous version would require a large research project.

One complexity is that the brain is an open system, in what would be formally called a non-equilibrium steady state, which continually receives new inputs and performs many computations at the same time, even though its entropy does not change that much over time. Landauer's principle, though, applies to drops in entropy that occur in each step of a calculation. Various other caveats would also be necessary. For example, there are long-range correlations between bits, and there are multiple heat baths in the brain.

As a simplified toy model, however, we can imagine that the brain computes in a serial fashion. It gets new inputs for each computation (thereby reinflating the entropy), and each computation causes a drop in entropy. In this case, the upper bound on bit-erasures suggested by Mr. Carlsmith would apply.

Prof. Wolpert's thinks that this calculation is legitimate as a first-pass, back-of-the-envelope upper bound on the bit-erasures that the brain could be implementing. It couldn't get published in a physics journal, but it might get published in a popular science journal, and it

helps get the conversation started. At a minimum, it's a strong concern that advocates of extreme amounts of computational complexity in the brain (for example, advocates of the view that you need much more than 1e22 FLOP/s to replicate the brain's computation) would need to address.

Adjustments to the estimate

Prof. Wolpert also expects that using Landauer's principle to estimate the amount of computation performed by the brain will result in substantial overestimates. A single neuron uses very complicated physical machinery to propagate a single bit along an axon. Prof. Wolpert expects this to be very far away from theoretical limits of efficiency.

That said, some computational processes in biology are very energy efficient. For example, Prof. Wolpert recently co-authored a paper on protein synthesis in ribosomes, showing that the energy efficiency of the computation is only around two orders of magnitude worse than Landauer's bound. Prof. Wolpert expects neurons to be much less efficient than this, but he doesn't know.

Emulating the brain (as opposed to just re-implementing its computation) could also introduce substantially additional compute costs, because the brain can implement operations directly, via its biochemistry, that could be difficult to simulate in silicon. And silicon computers are themselves seven or eight orders of magnitude less efficient than the theoretical limit set by Landauer's bound.