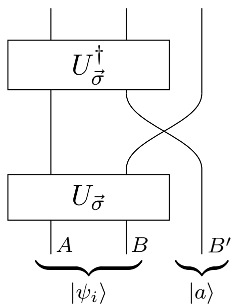
A Quantum Autoencoder: Diagrammatically

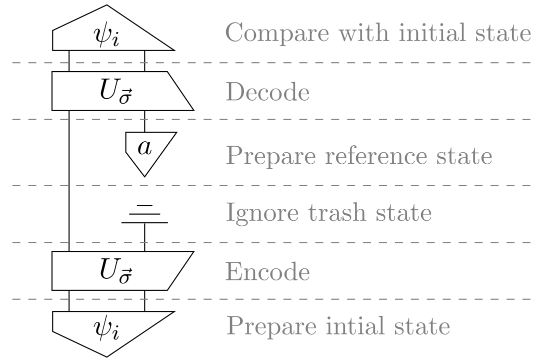
Suppose we have registers A, B and B' of n, k and k qubits respecively. In [1] a circuit is proposed for compressing an $(n+k)$-qubit state $|\psi_i\rangle$ on registers A and B into an n-qubit state, then (approximately) recovering it again. This is done by first preparing a blank 'reference' state $|a\rangle$ on B', then transforming $|\psi_i\rangle$ through a unitary $U_{\vec{\sigma}}$, indexed by parameters $\vec{\sigma} = (\sigma_1, \sigma_2, ...)$. To decode, the adjoint $U_{\vec{\sigma}}^{\dagger}$ is applied across registers A and B'. This circuit is drawn below:



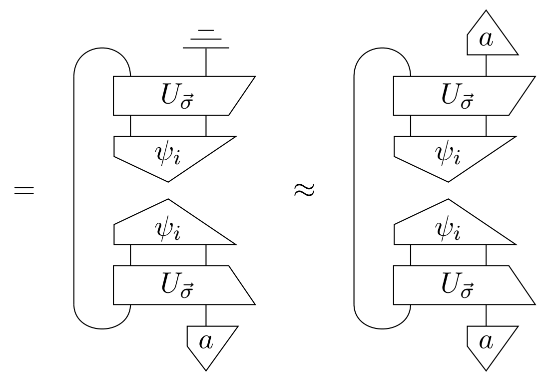
Given a mixture $\{p_i, |\psi_i\rangle\}$ of such $(n+k)$-qubits, the aim is to train a system to find the best possible encoder/decoder $U_{\vec{\sigma}}$ across all input states. This requires a cost function, the most obvious choice being the average fidelity $F(|\psi_i\rangle, \rho) = \langle\psi_i| \rho |\psi_i\rangle$ between the input state $|\psi_i\rangle$ and the recovered state ρ across all input states. Approximating this cost function would mean applying a SWAP test to the recovered state and a copy of the input state. This is relatively expensive, and requires the ability to prepare copies of the input state.

Instead, it is shown in [1] that - surprisingly - another good cost function is given by the average fidelity $F(|a\rangle, \rho')$ between the reference state $|a\rangle$ and the 'waste' state ρ'. The 'waste' state is the state of register B after the encoding unitary has been applied: $\rho' = \mathrm{Tr}_A[U_{\vec{\sigma}} |\psi_i\rangle \langle\psi_i| U_{\vec{\sigma}}^{\dagger}]$. Performing a SWAP test on these two states is cheaper (only a single unitary $U_{\vec{\sigma}}$ is applied) and requires no knowledge of $|\psi_i\rangle$.

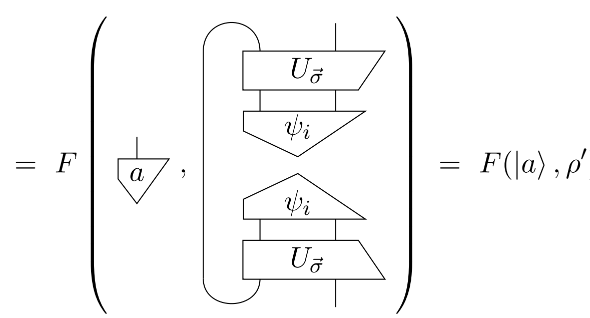
However, a diagrammatic approach reveals that this is perhaps not so surprising after all. We use the rigourous diagrammatic language set out in [2], where wires and boxes represent complex Hilbert spaces and linear maps respectively. Reading bottom-to-top, the 'obvious' thing to do is:



Now we're just looking for a simpler way to interpret this diagram. Since such diagrams are equal up to planar isotopoy, we can bend the top half around to the bottom to reveal that this diagram involves a trace over the first system. This can then be approximated by replacing the 'discard' effect $= \frac{1}{2^k}(1 \quad 1 \ ... \ 1)$ with $\langle a| = (a_1 \ a_2 \ ... \ a_k)$:



But then this is exactly:



i.e. the fidelity of the reference state and the waste state. So something that at first glance looked unobvious and counter-intuitive is in fact only a step or two away from being the most obvious thing possible.

[1] Romero, Jonathan, Jonathan P. Olson, and Alan Aspuru-Guzik. "Quantum autoencoders for efficient compression of quantum data" https://arxiv.org/pdf/1612.02806.pdf

[2] Coecke, Bob, and Aleks Kissinger. "Picturing quantum processes"