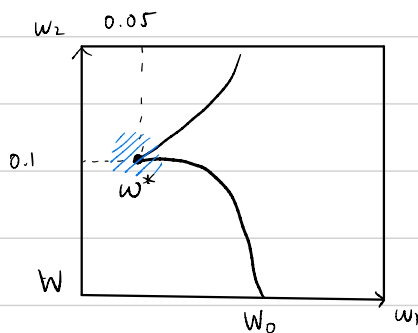


# SLT for Alignment 1


AI alignment is the problem of getting an AI system, potentially more intelligent and capable than any individual human or human civilisation altogether, to do what we want. To this end it seems robustly useful to understand the structure of knowledge and computations contained within neural networks. This will not solve Alignment on its own, but it seems hard to imagine good outcomes that do not involve radical progress on this front (for some definition of "understand" which might appear in a sentence like GPT-Safe understands GPT-4 understands GPT-5 ...).

## The high level principles behind our approach

1. Knowledge is a Singularity : we assume that the structure of knowledge and computations encoded in a trained neural network is reflected in the structure of the corresponding singularity. It is unclear to what extent this is true.



Knowledge  $w^*$ ,  $p(x|w^*)$

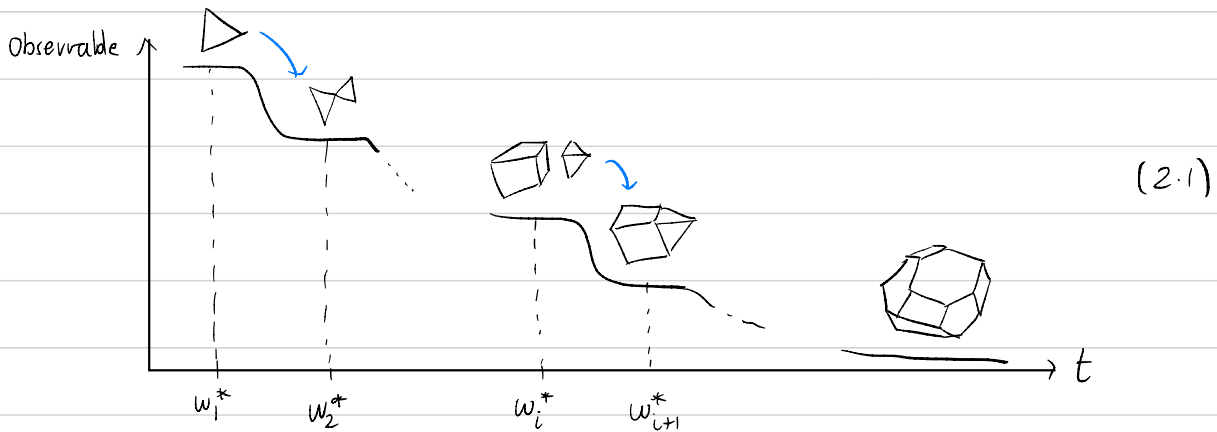
structure  $(K, w^*)$ , , ... (1.1)  
 $D_{\text{sing}}^b(K, w^*)$ , ...

2. Structure is constructed over Training : we assume that the structure of the final trained weight  $w^*$  is a function of the structures present in special intermediate states of training  $w_1^*, \dots, w_n^*$  which we call critical weights, and transformations between these structures.

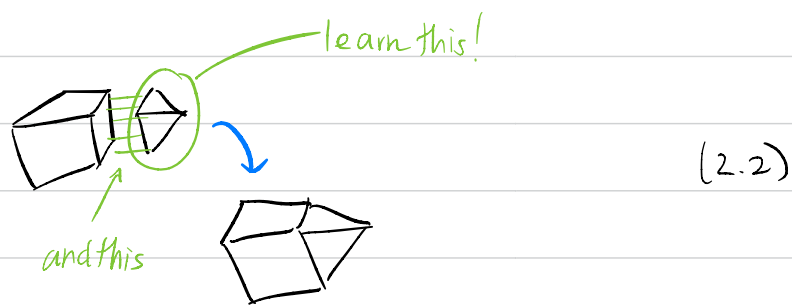
weight  $w_1^*$   $w_2^*$   $\dots$   $w_i^*$   $w_{i+1}^*$   $\dots$   $w^*$  (1.2)



3. Critical weights from Phase Transitions : we assume that the important changes in structure are associated with phase transitions (of the Bayesian posterior for example) and that these transitions can be detected cheaply during training, by developing and looking for divergences in generalised susceptibilities.



4. Structural inference is scalable we assume that given a critical weight  $w_i^*$  that the relevant change in structure to  $w_{i+1}^*$  can be effectively inferred from analysis of the model near  $w_i^*$ ,  $w_{i+1}^*$ . We call a device which performs this (and other structural inference) a spectroscopic probe by analogy with devices probing band structure of materials in solid state physics.



This probably depends on what we call the Locality Hypothesis : it is possible to choose the critical weights such that the transitions  $w_i^* \rightarrow w_{i+1}^*$  involve only a (relatively) small number of directions in weight space, or at least that structure decomposes in such a way.

5. Structure is Interpretable : we assume that the structures thereby inferred, and their transformations, serve usefully as "signs" for reasoning about and interpreting the knowledge and computations encoded in the network. That is, we assume

- Some structures can be put in correspondence with human recognisable concepts, logical rules, etc., by for example tracking phase transitions in performance in evaluations and associating this with results from structural inference.
- The signs and their relations can be used to understand circuits and other internal structures. Note the relation between structure in the above sense and "network structure" in the sense of architecture is nontrivial.

This seems to depend to some degree on the Universality Hypothesis which asserts that independently trained networks (on the same data) will discover similar structures, and moreover that these structures are similar to the ones humans see in the data. But there is no fundamental obstacle to using ML on top of the structural inference.

