

directions and then removing them to make future features easier to identify. But with superposition, one can't simply remove a direction even if one knows that it is a feature direction. [↗]

20. More formally, given a matrix $H \sim [d, m] = [h_0, h_1, \dots]$ of hidden layer activations $h \sim [m]$ sampled over d stimuli, if we believe there are n underlying features, we can try to find matrices $A \sim [d, n]$ and $B \sim [n, m]$ such that A is sparse. [↗]
21. In particular, it seems like we should expect to be able to reduce superposition at least a little bit with essentially no effect on performance, just by doing something like L1 regularization without any architectural changes. Note that models should have a level of superposition where the derivative of loss with respect to the amount of superposition is zero – otherwise, they'd use more or less superposition. As a result, there should be at least some margin within which we can reduce the amount of superposition without affecting model performance. [↗]
22. A more subtle issue is that GANs and VAEs often assume that their latent space is Gaussianly distributed. Sparse latent variables are very non-Gaussian, but central limit theorem means that the superposition of many such variables will gradually look more Gaussian. So the latent spaces of some generative models may in fact force models to use superposition! [↗]
23. Note that this has a nice information-theoretic interpretation: $\log(1 - S)$ is the surprisal of a given dimension being non-zero, and is multiplied by the expected number of non-zeros. [↗]
24. Note that in the compressed sensing case, the phase transition is in the limit as the number of dimensions becomes large – for finite-dimensional spaces, the transition is fast but not discontinuous. [↗]
25. We haven't encountered a specific term in the distributed coding literature that corresponds to this hypothesis specifically, although the idea of a "direction in activation-space" is common in the literature, which may be due to ignorance on our part. We call this hypothesis *linearity* [↗]
26. Experimental evidence seems to support this [55] [↗]
27. A related, but different, concept in the neuroscience literature is the "binding problem" [56] in which e.g. a red triangle is a co-occurrence of exactly one shape and exactly one color, which is not a representational challenge, but a binding problem arises if a decomposed code needs to represent simultaneously also a blue square — which shape feature goes with which color feature? Our work does not engage with the binding question, merely treating this as a co-occurrence of "blue", "red", "triangle", and "square". [↗]