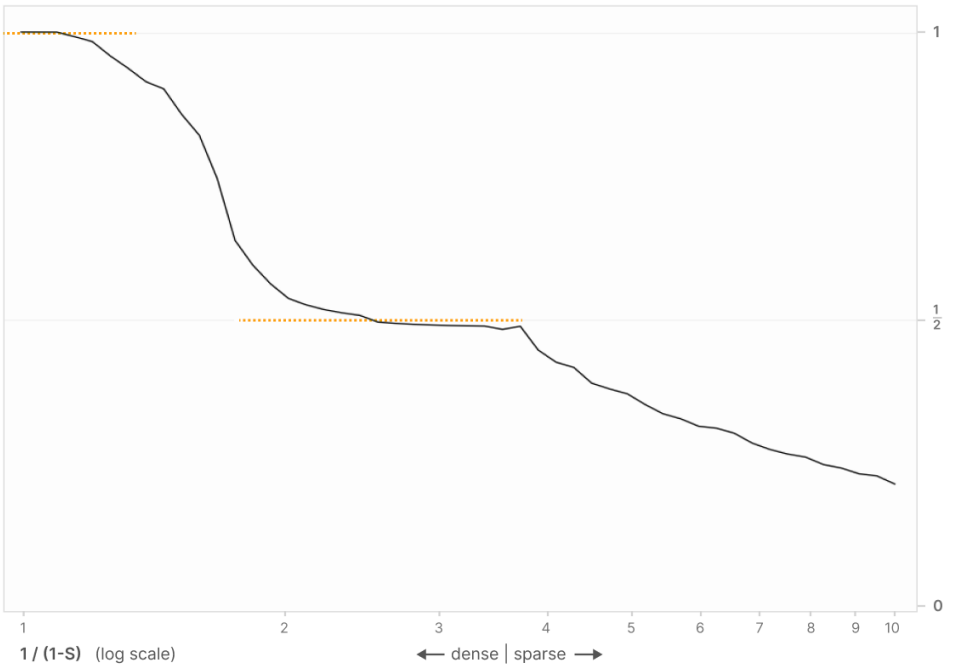
Uniform Superposition

As mentioned above, we begin our investigation with uniform superposition, where all features have the same importance and sparsity. We'll see later that this case has some unexpected structure, but there's also a much more basic reason to study it: it's much easier to reason about than the non-uniform case, and has fewer variables we need to worry about in our experiments.

We'd like to understand what happens as we change feature sparsity, S. Since all features are equally important, we will assume without loss of generality[14] that each feature has importance $I_i = 1$. We'll study a model with $n = 400$ features and $m = 30$ hidden dimensions, but it turns out the number of features and hidden dimensions doesn't matter very much. In particular, it turns out that the number of input features n doesn't matter as long as it's much larger than the number of hidden dimensions, $n \gg m$. And it also turns out that the number of hidden dimensions doesn't really matter as long as we're interested in the ratio of features learned to hidden features. Doubling the number of hidden dimensions just doubles the number of features the model learns.

A convenient way to measure the number of features the model has learned is to look at the Frobenius norm, $||W||_F^2$. Since $||W_i||^2 \simeq 1$ if a feature is represented and $||W_i||^2 \simeq 0$ if it is not, this is roughly the number of features the model has learned to represent. Conveniently, this norm is basis-independent, so it still behaves nicely in the dense regime $S = 0$ where the feature basis isn't privileged by anything and the model represents features with arbitrary directions instead.

We'll plot $D^* = m/||W||_F^2$, which we can think of as the "dimensions per feature":



Number of Hidden Dimensions per Embedded Feature

$m/||W||_F^2$

Unexpectedly, we find that the number of dimensions per feature is "sticky" at 1 and 1/2. We denote this with ····· .

Surprisingly, we find that this graph is "sticky" at 1 and $1/2$. (This very vaguely resembles the fractional quantum Hall effect – see e.g. this diagram.) Why is this? On inspection, the $1/2$ "sticky point" seems to correspond to a precise geometric arrangement where features come in "antipodal pairs", each being exactly the negative of the other, allowing two features to be packed into each hidden dimension. It appears that antipodal pairs are so effective that the model preferentially uses them over a wide range of the sparsity regime.

It turns out that antipodal pairs are just the tip of the iceberg. Hiding underneath this curve are a number of extremely specific geometric configurations of features.