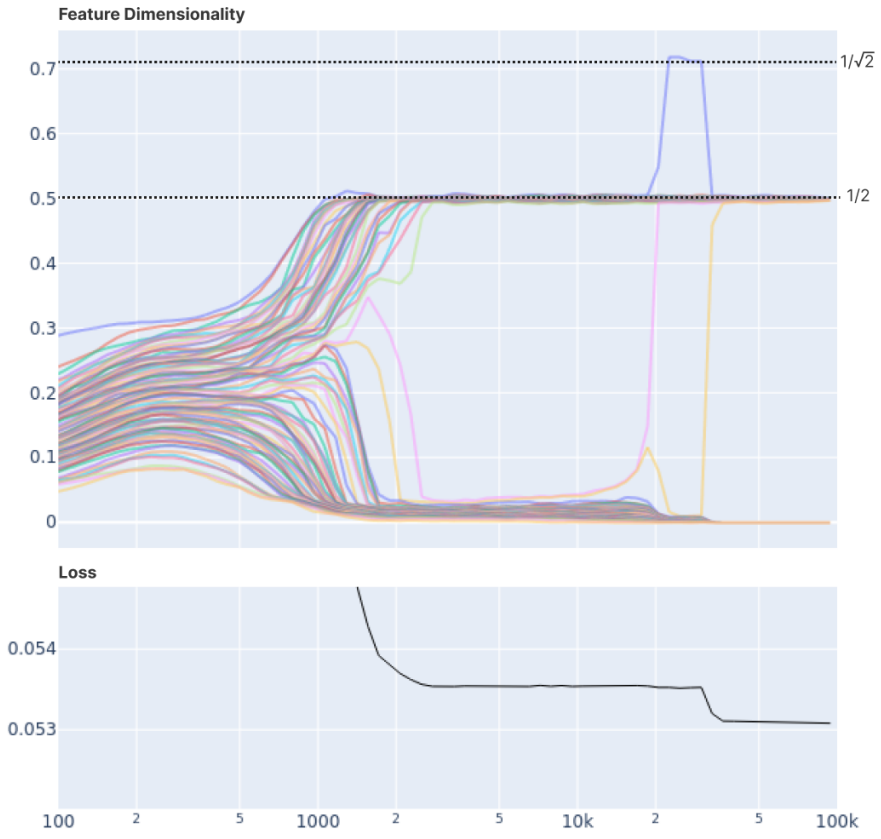

Note how the dimensionality of some features "jump" between different values and swap places. As this happens, the loss curve also undergoes a sudden drop (a very small one at the first jump, and a larger one at the second jump).

These results make us suspect that seemingly smooth decreases of the loss curve in larger models are in fact composed of many small jumps of features between different configurations. (For similar results of sudden mechanistic changes, see Olsson *et al.*'s induction head phase change [27], and Nanda and Lieberum's results on phase changes in modular arithmetic [28]. More broadly, consider the phenomenon of grokking [29].)

PHENOMENON 2: LEARNING AS GEOMETRIC TRANSFORMATIONS

Many of our toy model solutions can be understood as corresponding to geometric structures. This is especially easy to see and study when there are only $m = 3$ hidden dimensions, since we can just directly visualize the feature embeddings as points in 3D space forming a polyhedron.

It turns out that, at least in some cases, the learning dynamics leading to these structures can be understood as a sequence of simple, independent geometric transformations!

One particularly interesting example of this phenomenon occurs in the context of correlated features, as studied in the previous section. Consider the problem of representing $n = 6$ features in superposition within $m = 3$ dimensions. If we have the 6 features be 2 sets of 3 correlated features, we observe a really interesting pattern. The learning proceeds in distinct regimes which are visible in the loss curve, with each regime corresponding to a distinct geometric transformation: