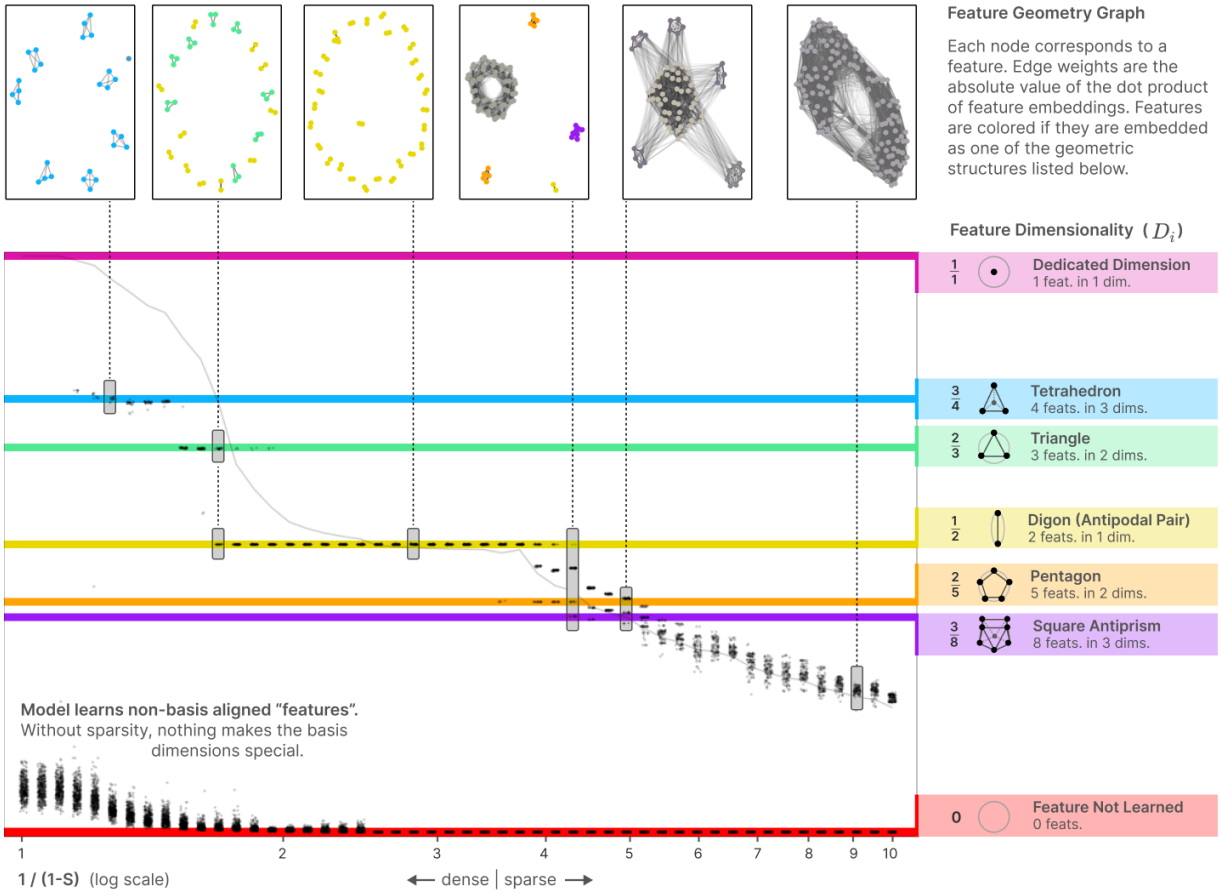


What is going on with the points clustering at specific fractions?? We'll see shortly that the model likes to create specific weight geometries and kind of jumps between the different configurations.

In the previous section, we developed a theory of superposition as a phase change. But everything on this plot between 0 (not learning a feature) and 1 (dedicating a dimension to a feature) is superposition. Superposition is what happens when features have fractional dimensionality. That is to say – superposition isn't just one thing!

How can we relate this to our original understanding of the phase change? We often think of water as only having three phases: ice, water and steam. But this is a simplification: there are actually many phases of ice, often corresponding to different crystal structures (eg. hexagonal vs cubic ice). In a vaguely similar way, neural network features seem to also have many other phases within the general category of "superposition."

WHY THESE GEOMETRIC STRUCTURES?

In the previous diagram, we found that there are distinct lines corresponding to dimensionality of: ¾ (tetrahedron), ⅔ (triangle), ½ (antipodal pair), ⅖ (pentagon), ⅜ (square antiprism), and 0 (feature not learned). We believe there would also be a 1 (dedicated dimension for a feature) line if not for the fact that basis features are indistinguishable from other directions in the dense regime.

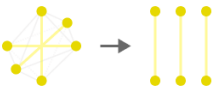
Several of these configurations may jump out as solutions to the famous Thomson problem. (In particular, square antiprisms are much less famous than cubes and are primarily of note for their role in molecular geometry due to being a Thomson problem solution.) As we saw earlier, there is a very real sense in which our model can be understood as solving a generalized version of the Thomson problem. When our model chooses to represent a feature, the feature is embedded as a point on an m-dimensional sphere.

A second clue as to what's going on is that there are lines for the Thomson solutions which are uniform polyhedra (e.g. tetrahedron), but there seem to be split lines where we'd expect to see non-uniform solutions (e.g. instead of a ⅗ line for triangular bipyramids we see a co-occurence of points at ⅔ for triangles and points at ½ for a antipodes). In a uniform polyhedron, all vertices have the same geometry, and so if we embed features as them each feature has the same dimensionality. But if we embed features as a non-uniform polyhedron, different features will have more or less interference with others.

In particular, many of the Thomson solutions can be understood as tegum products (an operation which constructs polytopes by embedding two polytopes in orthogonal subspaces) of smaller uniform polytopes. (In the earlier graph visualizations of feature geometry, two subgraphs are disconnected if and only if they are in different tegum factors.) As a result, we should expect their dimensionality to actually correspond to the underlying factor uniform polytopes.

A triangular bipyramid is the tegum product of a triangle and an antipode. As a result, we observe 3×2/3 features and 2×1/2 features, rather than 6×3/5 feaurs.

A pentagonal bipyramid is the tegum product of a pentagon and an antipode. As a result, we observe 5×2/5 features and 2×1/2 features, rather than 7×3/7 features.



An octahedron is the tegum product of three antipodes. This doesn't change the observed lines since 3/6=1/2.

This also suggests a possible reason why we observe 3D Thomson problem solutions, despite the fact that we're actually studying a higher dimensional version of the problem. Just as many 3D Thomson solutions are tegum products of 2D and 1D solutions, perhaps higher dimensional solutions are often tegum products of 1D, 2D, and 3D solutions.

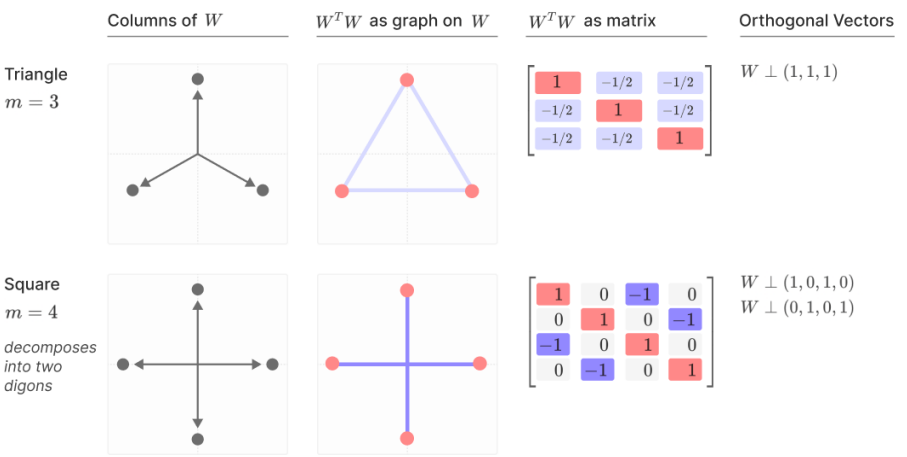
The orthogonality of factors in tegum products has interesting implications. For the purposes of superposition, it means that there can't be any "interference" across tegum-factors. This may be preferred by the toy model: having many features interfere simultaneously could be really bad for it. (See related discussion in our earlier mathematical analysis.)

Aside: Polytopes and Low-Rank Matrices

At this point, it's worth making explicit that there's a correspondence between *polytopes* and *symmetric, positive-definite, low-rank matrices* (i.e. matrices of the form $W^T W$). This correspondence underlies the results we saw in the previous section, and is generally useful for thinking about superposition.

In some ways, the correspondence is trivial. If one has a rank-m $n \times n$ -matrix of the form $W^T W$, then W is a $n \times m$ -matrix. We can interpret the columns of W as n points in a m-dimensional space. The place where this starts to become interesting is that it makes it clear that $W^T W$ is driven by the geometry. In particular, we can see how the off-diagonal terms are driven by the geometry of the points.

Put another way, there's an exact correspondence between polytopes and strategies for superposition. For example, every strategy for putting three features in superposition in a 2-dimensional space corresponds to a triangle, and every triangle corresponds to such a strategy. From this perspective, it doesn't seem surprising that if we have three equally important and equally sparse features, the optimal strategy is an equilateral triangle.



This correspondence also goes the other direction. Suppose we have a rank $(n-i)$ -matrix of the form W^TW. We can characterize it by the dimensions W *did not* represent – that is, which directions are orthogonal to W? For example, if we have a $(n-1)$ -matrix, we might ask what single direction did W not represent? This is especially informative if we assume that W^TW will be as "identity-like" as possible, given the constraint of not representing certain vectors.

In fact, given such a set of orthogonal vectors, we can construct a polytope by starting with n basis vectors and projecting them to a space orthogonal to the given vectors. For example, if we start in three dimensions and then project such that $W \perp (1,1,1)$, we get a triangle. More generally, setting $W \perp (1,1,1,...)$ gives us a regular n-simplex. This is interesting because it's in some sense the "minimal possible superposition." Assuming that features are equally important and sparse, the best possible direction to not represent is the fully dense vector $(1,1,1,...)$!

Non-Uniform Superposition

So far, this section has focused on the geometry of uniform superposition, where all features are of equal importance, equal sparsity, and independent. The model is essentially solving a variant of the Thomson problem. Because all features are the same, solutions corresponding to uniform polyhedra get especially low loss. In this subsection, we'll study non-uniform superposition, where features are somehow not uniform. They may vary in importance and sparsity, or have a correlational structure that makes them not independent. This distorts the uniform geometry we saw earlier.

In practice, it seems like superposition in real neural networks will be non-uniform, so developing an understanding of it seems important. Unfortunately, we're far from a comprehensive theory of the geometry of non-uniform superposition at this point. As a result, the goal of this section will merely be to highlight some of the more striking phenomena we observe:

- **Features varying in importance or sparsity** causes smooth deformation of polytopes as the imbalance builds, up until a critical breaking point at which they snap to another polytope.
- **Correlated features** prefer to be orthogonal, often forming in different tegum factors. As a result, correlated features may form an orthogonal local basis. When they can't be orthogonal, they prefer to be side-by-side. In some cases correlated features merge into a single feature: this hints at some kind of interaction between "superposition-like behavior" and "PCA-like behavior".
- **Anti-correlated features** prefer to be in the same tegum factor when superposition is necessary. They prefer to have negative interference, ideally being antipodal.

We attempt to illustrate these phenomena with some representative experiments below.