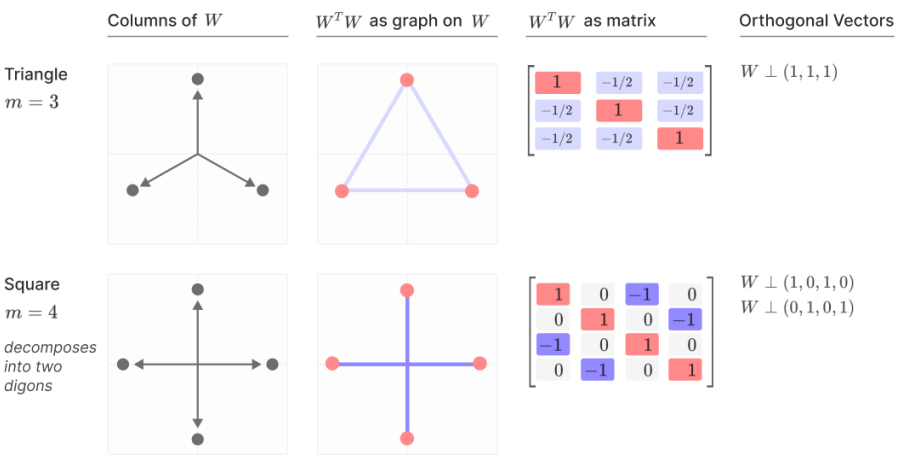

This correspondence also goes the other direction. Suppose we have a rank $(n-i)$ -matrix of the form W^TW. We can characterize it by the dimensions W *did not* represent – that is, which directions are orthogonal to W? For example, if we have a $(n-1)$ -matrix, we might ask what single direction did 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.