Sparsity-Relative Importance Phase Diagram (n=2, m=1)

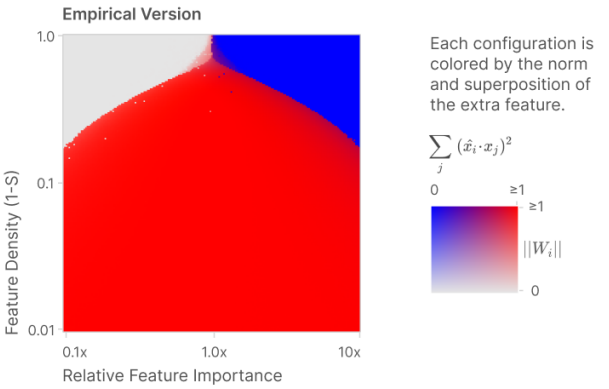
What happens to an "extra feature" if the model can't give each feature a dimension? There are three possibilities, depending on feature sparsity and the extra feature's importance relative to other features:

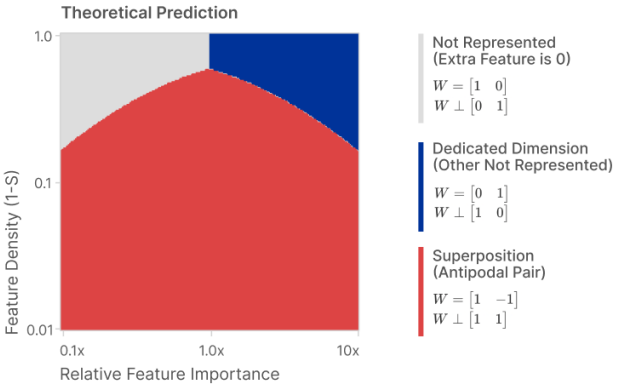
Extra Feature is Not Represented

Extra Feature Gets Dedicated Dimension

Extra Feature is Stored In Superposition

We can both study this empirically and build a theoretical model:





As expected, sparsity is necessary for superposition to occur, but we can see that it interacts in an interesting way with relative feature importance. But most interestingly, there appears to be a real phase change, observed in both the empirical and theoretical diagrams! The optimal weight configuration discontinuously changes in magnitude and superposition. (In the theoretical model, we can analytically confirm that there's a first-order phase change: there's crossover between the functions, causing a discontinuity in the derivative of the optimal loss.)

We can ask this same question of embedding three features in two dimensions. This problem still has a single "extra feature" (now the third one) we can study, asking what happens as we vary its importance relative to the other two and change sparsity.

For the theoretical model, we now consider four natural solutions. We can describe solutions by asking "what feature direction did W ignore?" For example, W might just not represent the extra feature – we'll write this $W \perp [0, 0, 1]$. Or W might ignore one of the other features, $W \perp [1, 0, 0]$. But the interesting thing is that there are two ways to use superposition to make antipodal pairs. We can put the "extra feature" in an antipodal pair with one of the others ($W \perp [0, 1, 1]$) or put the other two features in superposition and give the extra feature a dedicated dimension ($W \perp [1, 1, 0]$). Details on the closed form losses for these solutions can be found in this notebook. We do not consider a last solution of putting all the features in joint superposition, $W \perp [1, 1, 1]$.