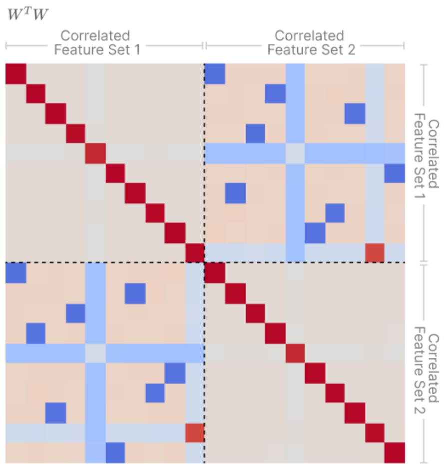


Models prefer to represent correlated features in orthogonal dimensions, creating "local orthogonal bases".

We train a model with 2 sets of 10 correlated features (n=20 total) with m=10 hidden dimensions.

Within each set of correlated featuers, the model creates a *local orthogonal basis*, having each feature be represented orthogonally.

Weight Element Values

-1 0 1

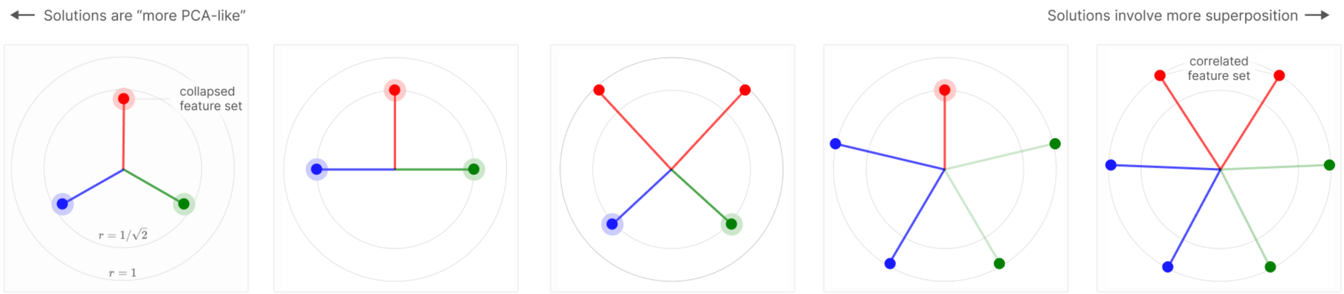
If this result holds in real neural networks, it suggests we might be able to make a kind of "local non-superposition" assumption, where for certain sub-distributions we can assume that the activating features are not in superposition. This could be a powerful result, allowing us to confidently use methods such as PCA which might not be principled to generally use in the context of superposition.

COLLAPSING OF CORRELATED FEATURES

One of the most interesting properties is that there seems to be a trade off with Principal Components Analysis (PCA) and superposition. If there are two correlated features a and b, but the model only has capacity to represent one, the model will represent their principal component $(a+b)/\sqrt{2}$, a sparse variable that has more impact on the loss than either individually, and ignore the second principal component $(a-b)/\sqrt{2}$.

As an experiment, we consider six features, organized into three sets of correlated pairs. Features in each correlated pair are represented by a given color (red, green, and blue). The correlation is created by having both features always activate together – they're either both zero or neither zero. (The exact non-zero values they take when they activate is uncorrelated.)

As we vary the sparsity of the features, we find that in the very sparse regime, we observe superposition as expected, with features arranged in a hexagon and correlated features side-by-side. As we decrease sparsity, the features progressively "collapse" into their principal components. In very dense regimes, the solution becomes equivalent to PCA.



Most PCA-like Solution
Approximately 0.5 ≤ 1-S

All Sets of Features Collapsed
Approximately 0.25 ≤ 1-S ≤ 0.5

Two Sets of Features Collapsed
Approximately 0.15 ≤ 1-S ≤ 0.2

One Set of Features Collapsed
Approximately 0.05 ≤ 1-S ≤ 0.15

No Features Collapsed
Approximately 1-S ≤ 0.05