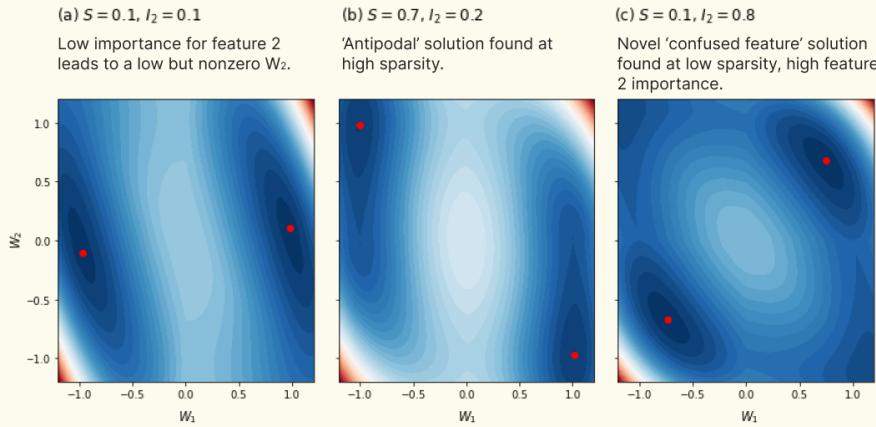


**Figure 1:** Loss surfaces for the  $n=2, m=1$  case of the ReLU output toy model.



Although many of these loss surfaces (Figure 1a, 1b) have minima qualitatively similar to one of the network weights used in the section [Superposition as a Phase Change](#), we also find a new phase where  $W_1 \simeq W_2 \simeq \frac{1}{\sqrt{2}}$ : weights are similar rather than antipodal. This 'confused feature' regime occurs when sparsity is low and both features are important (Figure 1c). (This is slightly similar to the behavior described in [The Geometry of Superposition – Collapsing of Correlated Features](#), but occurs without the features being correlated!) Further, although the solutions we find are often qualitatively similar to the weights used in [Superposition as a Phase Change](#), they can be quantitatively different, as Figure 1a shows. The transition from Figure 1a to Figure 1b is continuous: the minima moves smoothly in weight space as the degree of sparsity alters. This explains the 'blurry' region around the triple point in the phase diagram.

As Figure 1c shows, some combinations of sparsity and relative feature importance lead to loss surfaces with two minima (once the symmetry  $(W_1, W_2) \rightarrow (-W_1, -W_2)$  has been accounted for). If this pattern holds for larger values of  $n$  and  $m$  (and we see no reason why it would not) this could account for the [Discrete "Energy Level" Jumps phenomenon](#) as solutions hop between minima. In some cases (e.g. when parameters approach those needed for a phase transition) the global minimum can have a considerably smaller basin of attraction than local minima. The transition between the antipodal and confused-feature solutions appears to be discontinuous.

**Original Authors' Response:** This closed form analysis of the  $n = 2, m = 1$  case is fascinating. We hadn't realized that  $W_1 \simeq W_2 \simeq \frac{1}{\sqrt{2}}$  could be a solution without correlated features! The clarification of the "blurry behavior" and the observation about local minima are also very interesting. More generally, we're very grateful for the independent replication of our core results.