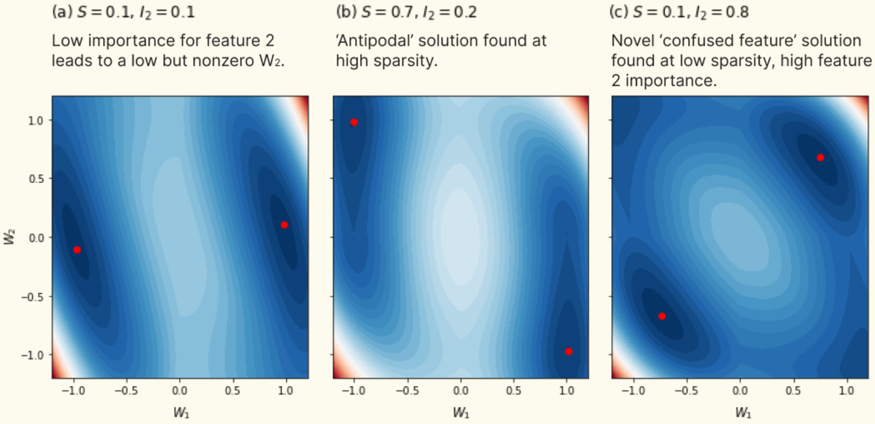
Figure 1: Loss surfaces for the n=2, m=1 case of the ReLU output toy model.



(a) $S = 0.1$, $I_2 = 0.1$

Low importance for feature 2 leads to a low but nonzero W₂.

(b) $S = 0.7$, $I_2 = 0.2$

'Antipodal' solution found at high sparsity.

(c) $S = 0.1$, $I_2 = 0.8$

Novel 'confused feature' solution found at low sparsity, high feature 2 importance.

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 $(W_1, W_2) \to (-W_1, -W_2)$ 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.

REPLICATION

__Jeffrey Wu__ and __Dan Mossing__ are members of the Alignment team at OpenAI.

We are very excited about these toy models of polysemanticity. This work sits at a rare intersection of being plausibly very important for training more interpretable models and being very simple and elegant. The results have been surprisingly easy to replicate -- we have reproduced (with very little fuss) plots similar to those in the Demonstrating Superposition – Basic Results, Geometry – Feature Dimensionality, and Learning Dynamics – Discrete "Energy Level" Jumps sections.

Original Authors' Response: We really appreciate this replication of our basic results. Some of our findings were quite surprising to us, and this gives us more confidence that they aren't the result of an idiosyncratic quirk or bug in our implementations.

Code

We provide a notebook to reproduce some of the core diagrams in this article here. (It isn't comprehensive, since we needed to rewrite code for our experiments to run outside our codebase.) We provide a separate notebook for the theoretical phase change diagrams.

Note that the reproductions by other researchers mentioned in comments above were not based on this code, but are instead fully independent replications with clean code from the description in an early draft of this article.

Acknowledgments

We're extremely grateful to a number of colleagues across several organizations for their invaluable support in our writing of this paper.

Jeff Wu, Daniel Mossing, Tom McGrath, and Kshitij Sachan did independent replications of many of our experiments, greatly increasing our confidence in our results. Kshitij Sachan's and Tom McGrath's additional investigations and insightful questions both pushed us to clarify our understanding of the superposition phase change (both as reflected in this paper, and in further understanding which we learned from them not captured here). Buck Shlegeris, Adam Scherlis, and Adam Jermyn shared valuable insights into the mathematical nature of the toy problem and related work. Adam Jermyn also coined the term "virtual neurons."

Gabriel Goh, Neel Nanda, Vladimir Mikulik, and Nick Cammarata gave detailed feedback which improved the paper, in addition to being motivating. Alex Dimakis, Piotr Indyk, Dan Yamins generously took time to discuss these results with us and give advice on how they might connect to their area of expertise. Finally, we benefited from the feedback and comments of James Bradbury, Sebastian Farquhar, Shan Carter, Patrick Mineault, Alex Tamkin, Paul Christiano, Evan Hubinger, Ian McKenzie, and Sid Black.

Finally, we're very grateful to all our colleagues at Anthropic for their advice and support: Daniela Amodei, Jack Clark, Tom Brown, Ben Mann, Nick Joseph, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones,, Timothy Telleen-Lawton, Anna Chen, Yuntao Bai, Jeffrey Ladish, Deep Ganguli, Liane Lovitt, Nova DasSarma, Jia Yuan Loke, Jackson Kernion, Tom Conerly, Scott Johnston, Jamie Kerr, Sheer El Showk, Stanislav Fort, Rebecca Raible, Saurav Kadavath, Rune Kvist, Jarrah Bloomfield, Eli Tran-Johnson, Rob Gilson, Guro Khundadze, Filipe Dobreira, Ethan Perez, Sam Bowman, Sam Ringer, Sebastian Conybeare, Jeeyoon Hyun, Michael Sellitto, Jared Mueller, Joshua Landau, Cameron McKinnon, Sandipan Kundu, Jasmine Brazilek, Da Yan, Robin Larson, Noemí Mercado, Anna Goldie, Azalia Mirhoseini, Jennifer Zhou, Erick Galankin, James Sully, Dustin Li, James Landis.

Author Contributions

Basic Results - The basic toy model results demonstrating the existence of superposition were done by Nelson Elhage and Chris Olah. Chris suggested the toy model and Nelson ran the experiments.

Phase Change - Chris Olah ran the empirical phase change experiments, with help from Nelson Elhage. Martin Wattenberg introduced the theoretical model where exact losses for specific weight configurations can be computed.

Geometry - The uniform superposition geometry results were discovered by Nelson Elhage and Nicholas Schiefer, with help from Chris Olah. Nelson discovered the original $m/||W||_F^2$ mysterious "stickiness". Chris introduced the definition of feature dimensionality. Nicholas and Nelson then investigated the polytopes that formed. As for non-uniform superposition, Martin Wattenberg performed the initial investigations of the resulting geometry, focusing on the behavior of correlated features. Chris extended this with an investigation of the role of relative feature importance and sparsity.

Learning Dynamics - Nelson Elhage discovered the "energy level jump" phenomenon, in collaboration with Nicholas Schiefer and Chris Olah. Martin Wattenberg discovered the "geometric transformations" phenomenon.

Adversarial Examples - Chris Olah and Catherine Olsson found evidence of a connection between superposition and adversarial examples.

Superposition with a Privileged Basis / Doing Computation - Chris Olah did the basic investigation of superposition in a privileged basis. Nelson Elhage, with help from Chris, investigated the "absolute value" model which provided a more principled demonstration of superposition and showed that computation could be done while in superposition. Nelson discovered the "asymmetric superposition" motif.

Theory - The theoretical picture articulated over the course of this paper (especially in the "mathematical understanding" section) was developed in conversations between all authors, but especially Chris Olah, Jared Kaplan, Martin Wattenberg, Nelson Elhage, Tristan Hume, Tom Henighan, Catherine Olsson, Nicholas Schiefer, Dawn Drain, Shauna Kravec, Roger Grosse, Robert Lasenby, and Sam McCandlish. Jared introduced the strategy of rewriting the loss by grouping terms with the number of active features. Both Jared and Martin independently noticed the value of investigating the $n = 2; m = 1$ case as the simplest case to understand. Nicholas and Dawn clarified our understanding of the connection to compressed sensing.