- **Is there a statistical test for catching superposition?**
- **How can we control whether superposition and polysemanticity occur?** Put another way, can we change the phase diagram such that features don't fall into the superposition regime? Pragmatically, this seems like the most important question. L1 regularization of activations, adversarial training, and changing the activation function all seem promising.
- **Are there any models of superposition which have a closed-form solution?** Saxe et al. [26] demonstrate that it's possible to create nice closed-form solutions for linear neural networks. We made some progress towards this for the $n = 2; m = 1$ ReLU output model (and Tom McGrath makes further progress in his comment), but it would be nice to solve this more generally.
- **How realistic are these toy models?** To what extent do they capture the important properties of real models with respect to superposition? How can we tell?
- **Can we estimate the feature importance curve or feature sparsity curve of real models?** If one takes our toy models seriously, the most important properties for understanding the problem are the feature importance and sparsity curves. Is there a way we can estimate them for real models? (Likely, this would involve training models of varying sizes or amounts of regularization, observing the loss and neuron sparsities, and trying to infer something.)
- **Should we expect superposition to go away if we just scale enough?** What assumptions about the feature importance curve and sparsity would need to be true for that to be the case? Alternatively, should we expect superposition to remain a constant fraction of represented features, or even to increase as we scale?
- **Are we measuring the maximally principled things?** For example, what is the most principled definition of superposition / polysemanticity?
- **How important are polysemantic neurons?** If X% of the model is interpretable neurons and 1-X% are polysemantic, how much should we believe we understand from understanding the x% interpretable neurons? (See also the "feature packing principle" suggested above.)
- **How many features should we expect to be stored in superposition?** This was briefly discussed in the previous section. It seems like results from compressed sensing should be able to give us useful upper-bounds, but it would be nice to have a clearer understanding – and perhaps tighter bounds!
- **Does the apparent phase change we observe in features/neurons have any connection to phase changes in compressed sensing?**
- **How does superposition relate to non-robust features?** An interesting paper by Gabriel Goh (archive.org backup) explores features in a linear model in terms of the principal components of the data. It focuses on a trade off between "usefulness" and "robustness" in the principal component features, but it seems like one could also relate it to the interpretability of features. How much would this perspective change if one believed the superposition hypothesis – could it be that the useful, non-robust features are an artifact of superposition?
- **To what extent can neural networks "do useful computation" on features in superposition?** Is the absolute value problem representative of computation in superposition generally, or idiosyncratic? What class of computation is amenable to being performed in superposition? Does it require a sparse structure to the computation?
- **How does superposition change if features are not independent?** Can superposition pack features more efficiently if they are anti-correlated?
- **Can models effectively use nonlinear representations?** We suspect models will tend not to use them, but further experimentation could provide good evidence. See the appendix on nonlinear compression. For example investigating the representations used by autoencoders with multi-layer encoders and decoders with really small bottlenecks on random uncorrelated data.