A trivial problem, that is hard for any kernelizable algorithm

Random data matrix, labeled by one of the features

Train on subset of rows
- labeled with some target column
- loss averaged over all $n$ examples

Solution sparse & linear: unit vector $e_i$ picks out $i$th feature
Hardness for GD with 2-norm regularization

Provably hard for any algorithm predicting with $\hat{y} = \sigma(w \cdot x)$, where
- $w$ = linear combination of instances
- square, logistic, hinge loss
- *any embedding* of the instances
Conjecture

Problem remains **hard** for any **deep neural net** trained with Gradient Descent + 2-norm regularization

Adding **hidden layers** does not help
Changing **transfer function** does not help
**Dropout** does not help

**Only experimental evidence**

**1-norm regularization** works fine