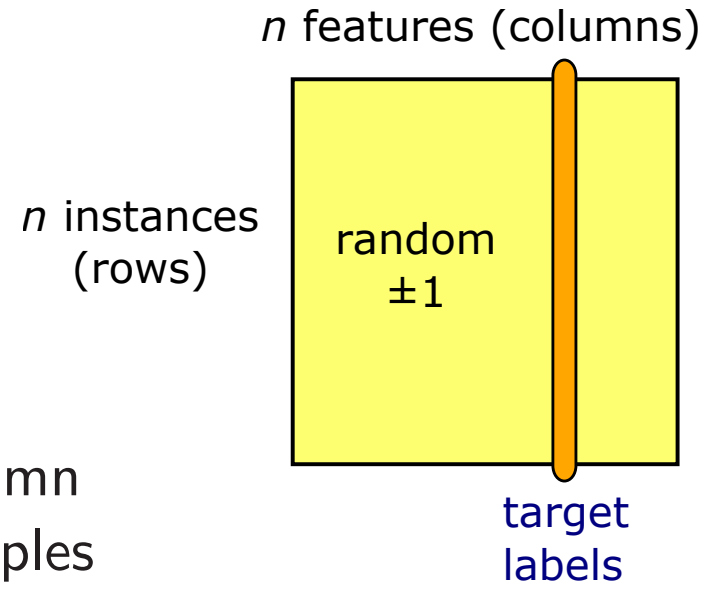
The limits of squared Euclidean distance regularization

Michał Dereziński (speaker), Manfred K. Warmuth

A trivial problem, that is hard for any **kernelizable algorithm**

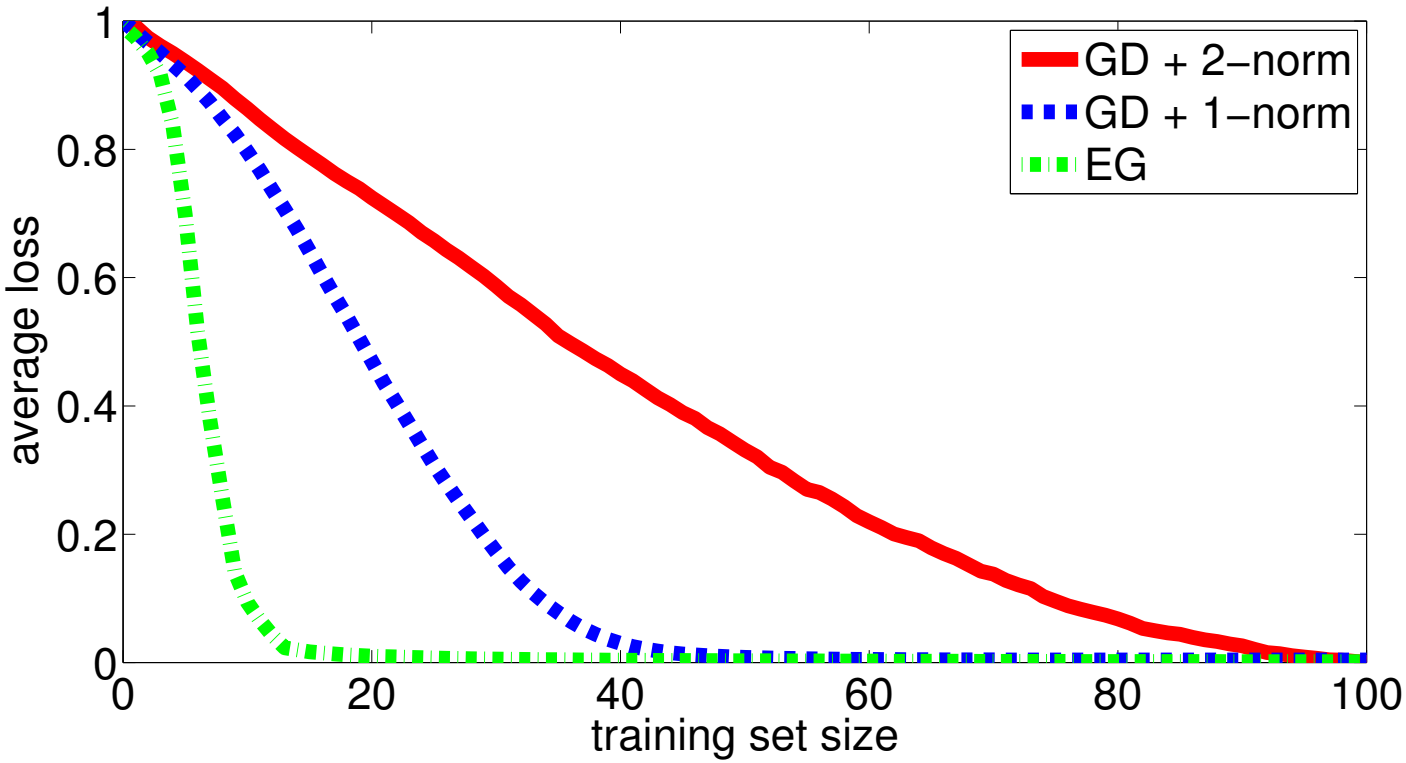
Random data matrix, labeled by one of the features



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

Solution sparse & linear: unit vector $\mathbf{e}_i$ picks out ith feature

Hardness for GD with 2-norm regularization



Provably hard for any algorithm predicting with $\widehat{y} = \sigma(\mathbf{w} \cdot \mathbf{x})$, where
- $\mathbf{w}$ = 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