



Random deep neural networks are biased towards simple functions

Giacomo De Palma Bobak Toussi Kiani Seth Lloyd

The generalization problem

- Deep neural networks do not overfit despite # weights \gg # training points
- Conjecture [Valle Pérez et al., ICLR 2019]: deep neural networks are biased towards simple functions, weights highly redundant
- Experiments on bit string classifiers generated by deep neural networks with randomly initialized weights: classifiers with low Lempel-Ziv complexity occur with higher probability

Deep neural networks as Gaussian processes

(Google Brain)

- Random weights with normal iid Gaussian distribution
- Limit of infinite width of hidden layers
- Generated function ϕ is Gaussian process: for any x_1, \dots, x_k in R^n , $\phi(x_1), \dots, \phi(x_k)$ have correlated Gaussian distribution

Binary classifiers of bit strings

- Encode strings of n bits in R^n as $\{-1, 1\}^n$
- Input x classified as $\text{sign}(\phi(x))$
- Hamming distance $h(x, y) = \#$ of different bits
- Uniformly random classifier: for any x in $\{-1, 1\}^n$ with high probability there exists y with different classification at $h(x, y) = 1$
- Same properties for classifiers generated by random deep neural networks?

Main results

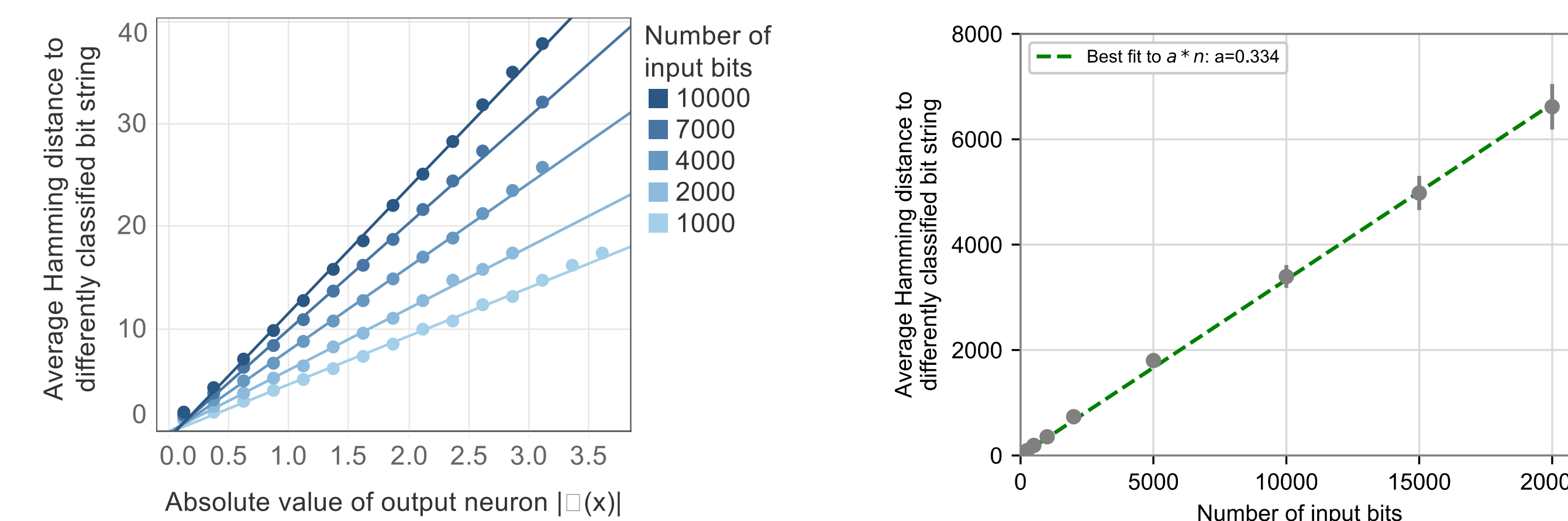
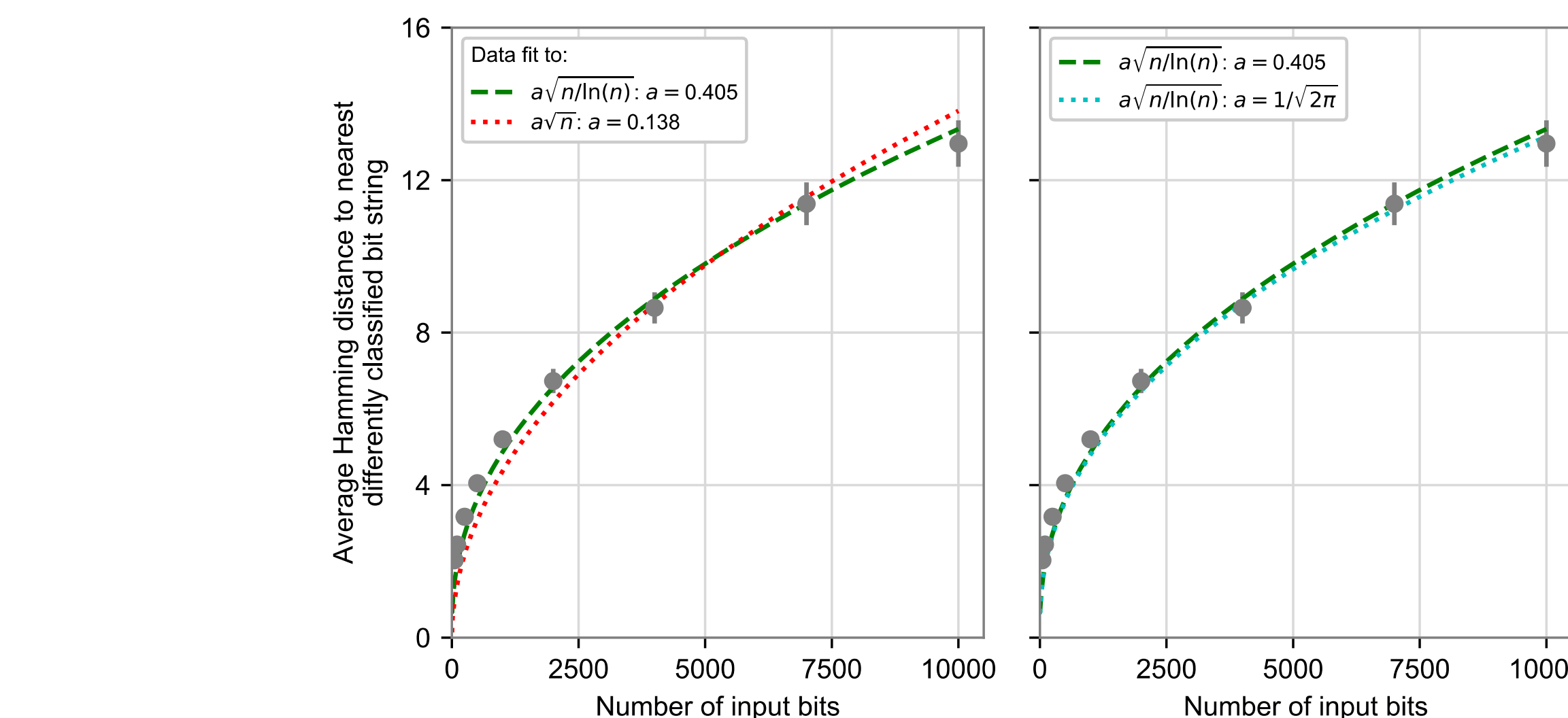
- For any input string x of n bits, the average Hamming distance of the closest bit string with a different classification is

$$\sqrt{\frac{n}{2\pi \ln n}}$$

and grows linearly with $\phi(x)$ as $\frac{|\phi(x)|}{2} \sqrt{\frac{n}{\ln n}}$

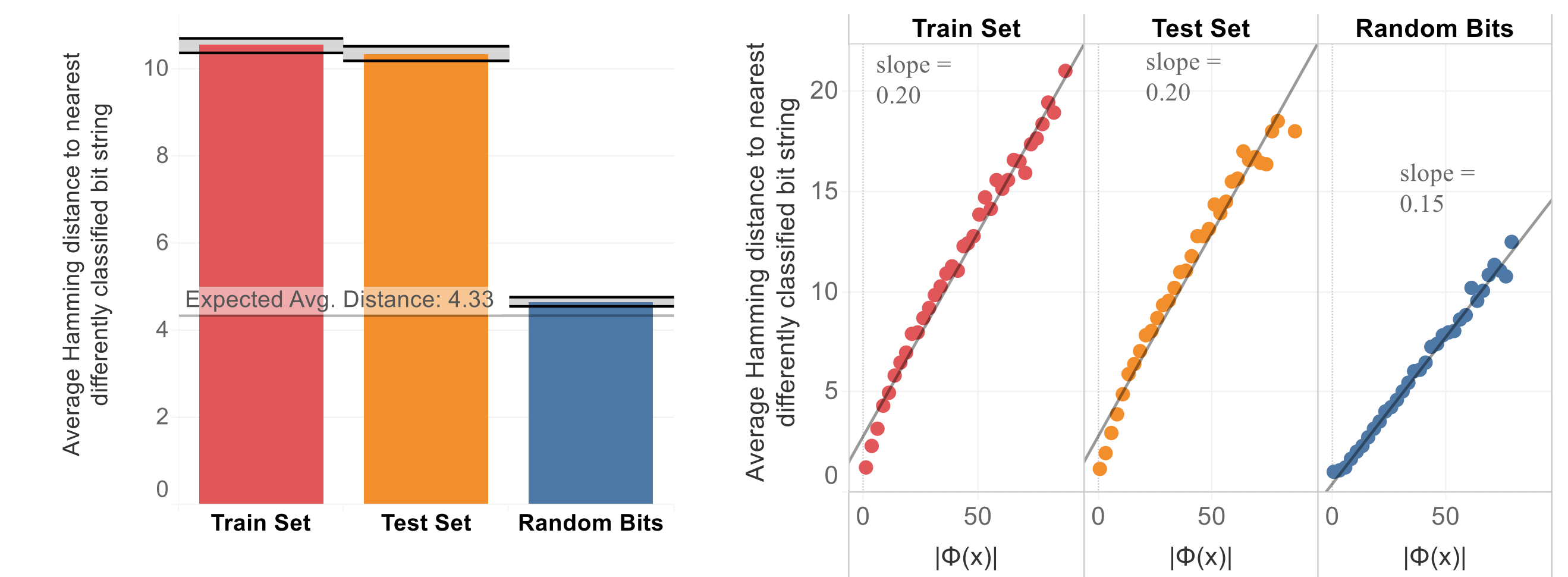
- For any input bit string x of n bits, the minimum number of random bit flips required to change the classification scales linearly with n

Experiments on random DNNs (2 hidden layers)



Experiments on MNIST (2 hidden layers)

- Robustness survives training



- Robustness correlated with generalization



Conclusions

- Binary classifiers of bit strings generated by random deep neural networks are biased towards simple functions
- Experiments: simplicity bias survives after training
- Analytical proof for trained DNNs??