Non-Vacuous Generalisation Bounds for Shallow Neural Networks

Felix Biggs[1,2] and Benjamin Guedj[1,2]

[1]University College London. [2]Inria London

Our Goal

Non-vacuous generalisation bounds for neural networks.

1. Train via standard SGD.
2. No randomisation of output (as in PAC-Bayes).

A partial solution

In single hidden layer networks with erf or GELU activations.

How?

Expressing the network as a PAC-Bayesian majority vote.

Basic PAC-Bayes Bound

- Randomised classification using $h \sim Q$.
- Misclassification loss.
- Fix a data-independent "prior", P over hypothesis space.

With probability $\geq 1-\delta$ over the sample for all Q,

$$\mathbb{E}_{h\sim Q}\left[L_{\text{out}}(h) - \hat{L}_{\text{in}}(h)\right] \leq O\left(\sqrt{\frac{\text{KL}(Q,P) + \log\frac{1}{\delta}}{\#\text{samples.}}}\right).$$

Majority Vote

MV_Q is the highest probability prediction under Q.

$$L(\mathrm{MV}_Q) \leq 2\, \mathbb{E}_{h\sim Q} L(h)$$

RHS is the quantity bounded by PAC-Bayes!

- Well-studied in PAC-Bayes literature.

Question

Can we construct Q with a neural network as its majority vote?

NNs as Majority Votes

For normalised data, consider networks

$$f(x) = V\phi(Ux).$$

Exact MV

Binary classification with $\phi = \mathrm{erf}$.

- Q randomised sign-activated network.

Approximate MV[1]

Multi-class with $\phi = \mathrm{erf}$ or $\phi = \mathrm{GELU}$.

- Q randomised network with more complex activations.

[1] Small additional assumption leads to bound.

	Data	Test Err	Bound	Data-dependent prior
ERF	Binary-MNIST	0.038	0.837	0.286
ERF	Binary-Fashion	0.085	0.426	0.297
ERF	MNIST	0.046	0.772	0.490
ERF	Fashion	0.150	0.984	0.727
GELU	MNIST	0.043	0.693	0.293
GELU	Fashion	0.153	0.976	0.568

Exciting link of NNs to Majority Votes.

Future Directions

1. Application to more complex network structures.
2. Use of more sophisticated majority vote tools.

Thanks for listening!