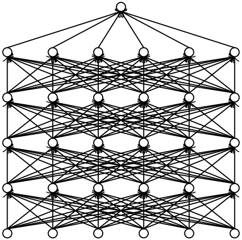
Generalization Bounds for Neural Networks via Approximate Description Length

Amit Daniely (Hebrew University and Google) and *Elad Granot (Hebrew University)*

December 1, 2019

What is the sample complexity of $\mathcal{N} = \{W_t \circ \rho \ldots \circ \rho \circ W_1\}$?



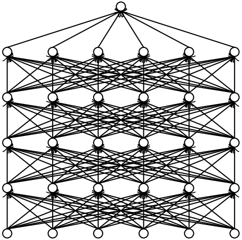
Input Space: $[-1,1]^d$
Width: d
Depth: t
Activation: ρ
Weights: $W_1, \ldots, W_{t-1} \in M_{d\times d}$ and $W_t \in M_{1,d}$

- Classical Theory: $\tilde{\Theta}(d^2)$ params $\Rightarrow \tilde{\Theta}(d^2)$ Sample Complexity

Why Networks Generalize with Much Fewer Exampels?

- Should we consider bounds on the weights?

$$\mathcal{N} = \{W_t \circ \rho \ldots \circ \rho \circ W_1 : \|W_i\|_F \leq R,\ \|W_i\| \leq 1\}$$



Input Space: $[-1,1]^d$
Width: d
Depth: t
Activation: ρ
Weights: $W_1, \ldots, W_{t-1} \in M_{d \times d}$ and $W_t \in M_{1,d}$

$$\mathcal{N} = \{W_t \circ \rho \ldots \circ \rho \circ W_1 : \|W_i\|_F \leq R, \ \|W_i\| \leq 1\}$$

- What is the sample complexity of $\mathcal{N}$?
- For linear $\rho(x) = x$, the sample complexity is $\tilde{\Theta}(dR^2)$
- Can we match this bound for non-linear ρ?
  - [Neyshabur, Tomioka, Srebro 15, Bartlett, Foster, Telgarsky 17, Neyshabur, Bhojanapalli, Srebro 18, Arora, Ge, Neyshabur, Zhang 18, Neyshabur, Li, Bhojanapalli, LeCun, Srebro 19,]: $\tilde{\Theta}(d^2R^2)$
  - This Work: Yes!

- Radamacher Complexity?
  - Talagrand's concentration lemma is loose in high dimension
  - Many experts failed
- A new technique: Approximate Description Length (ADL)
  - $\mathrm{ADL}(\mathcal{H})$: #bits required to approximately describe functions in $\mathcal{H}$
  - $\mathrm{ADL}(\mathcal{H}) = n \Rightarrow$ sample complexity $O\left(\frac{n}{\epsilon^2}\right)$
  - Develop tools to bound ADL
    - Get the correct value for linear classes
    - Behaves nicely with compositions (even in high dimension)