Complexity of Neural Networks

Kaifeng Bu

Mathematical Picture Language@ Harvard
Jan. 19 2021

[1] K. Bu, Y. Zhang, Q. Luo, arxiv 2010.07587

[2] K. Bu, D. Koh, L. Li, Q. Luo, Y. Zhang, arxiv 2101.06154

Outline

Part I: Complexity of functions

- **Introduction of neural networks**
- **Topological entropy**
- **Depth-width trade-off via topological entropy**

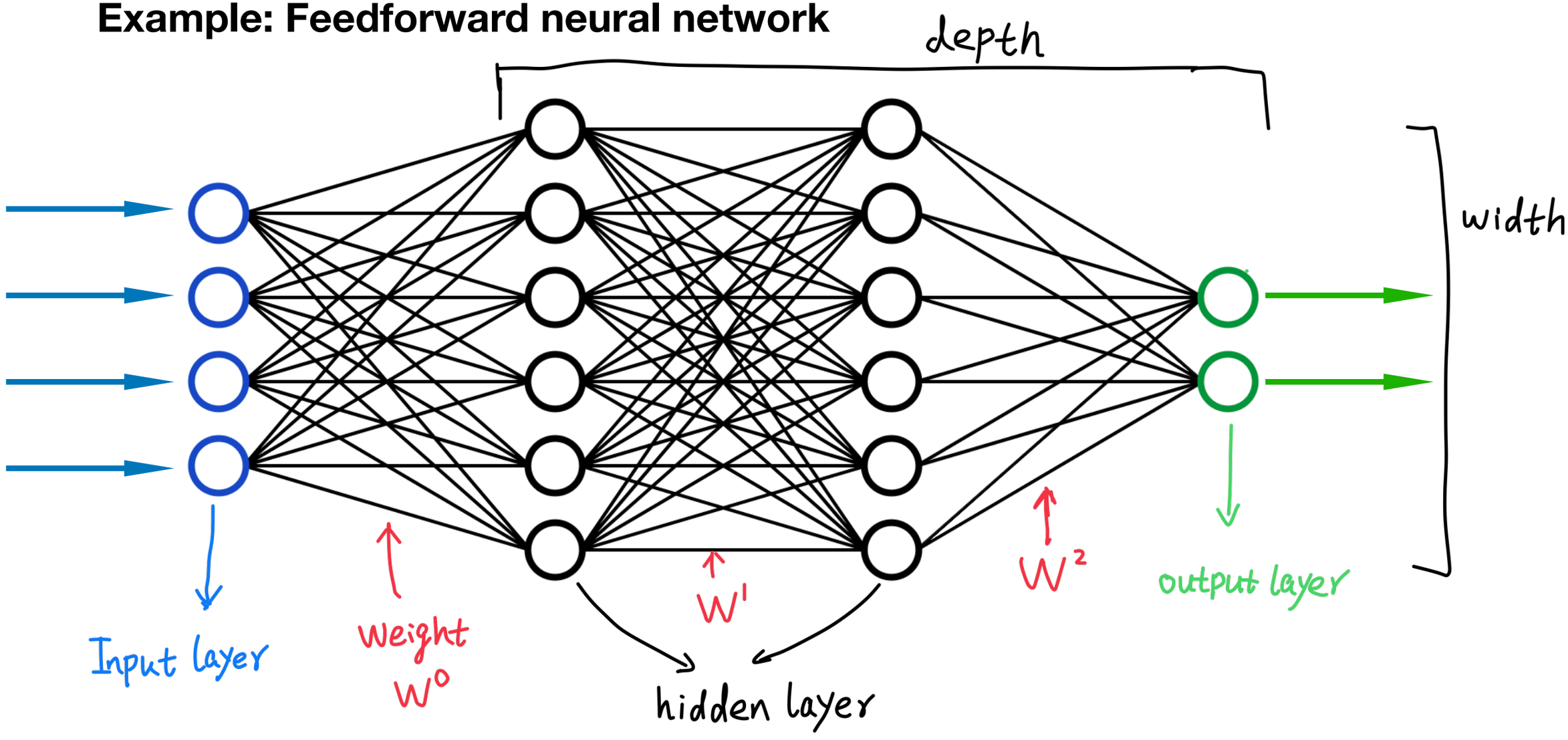
Part II: Complexity of function classes

- **Introduction of Rademacher complexity**
- **Rademacher complexity of quantum circuits**

Part III: Summary and further direction

Preliminary: Neural networks

Example: Feedforward neural network

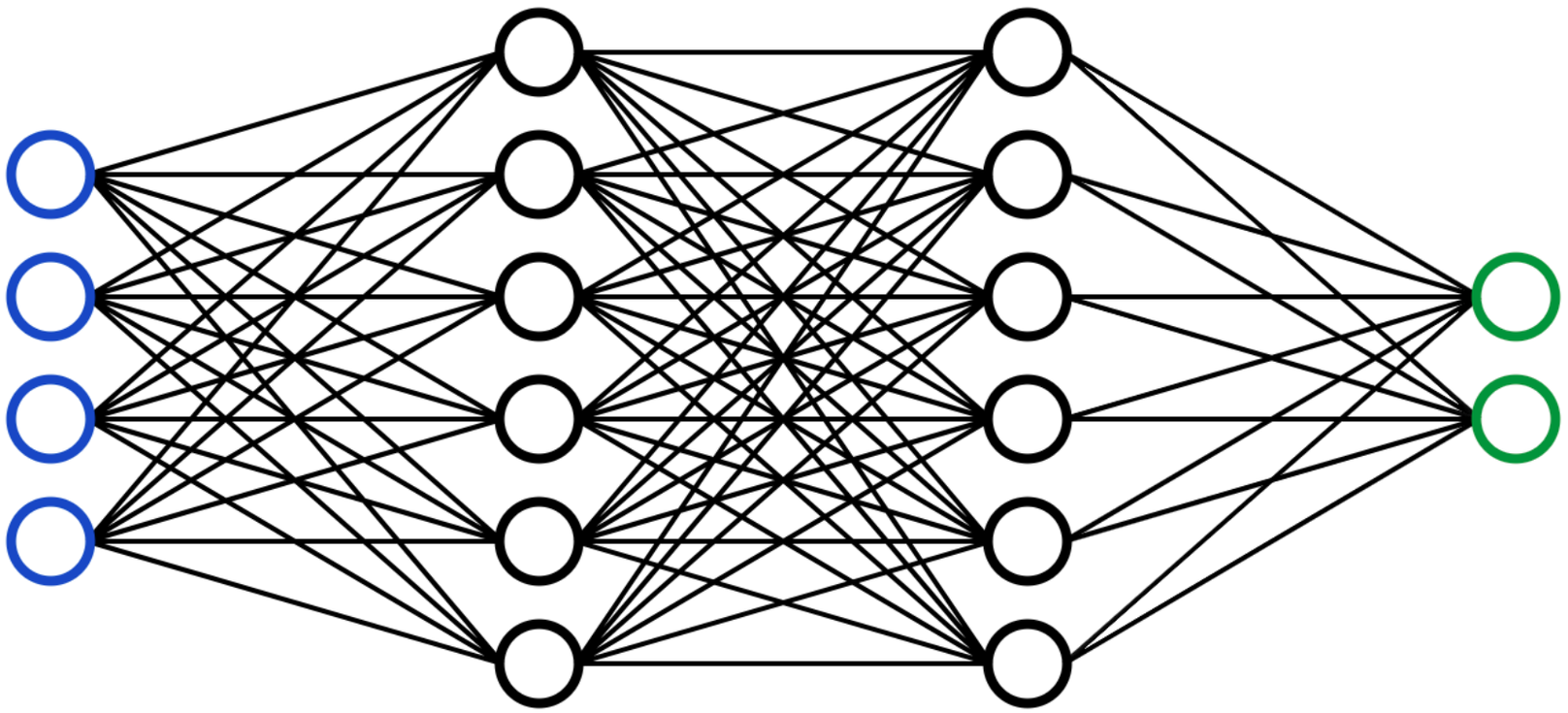


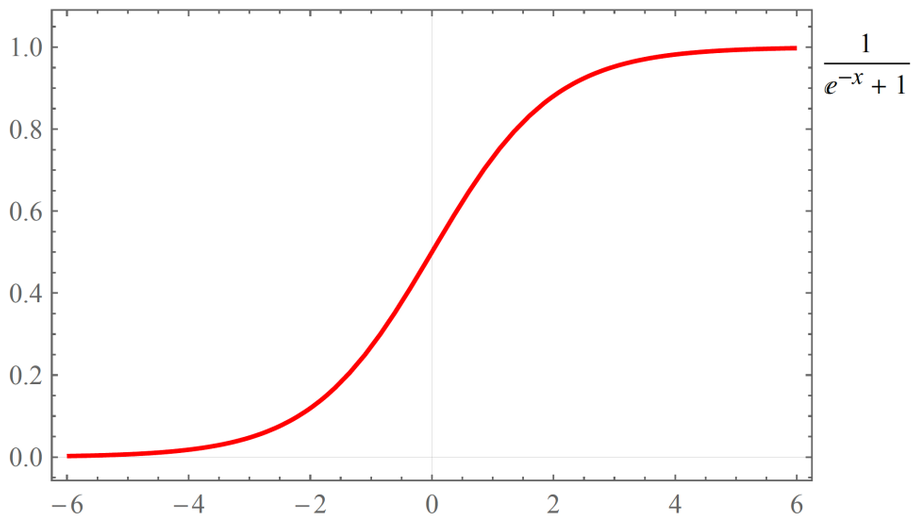
$$f(\vec{x}) = W_2\sigma(W_1\sigma(W_0\vec{x}))$$

Activation function $\sigma : \mathbb{R} \to \mathbb{R}$ (σ is nonlinear)

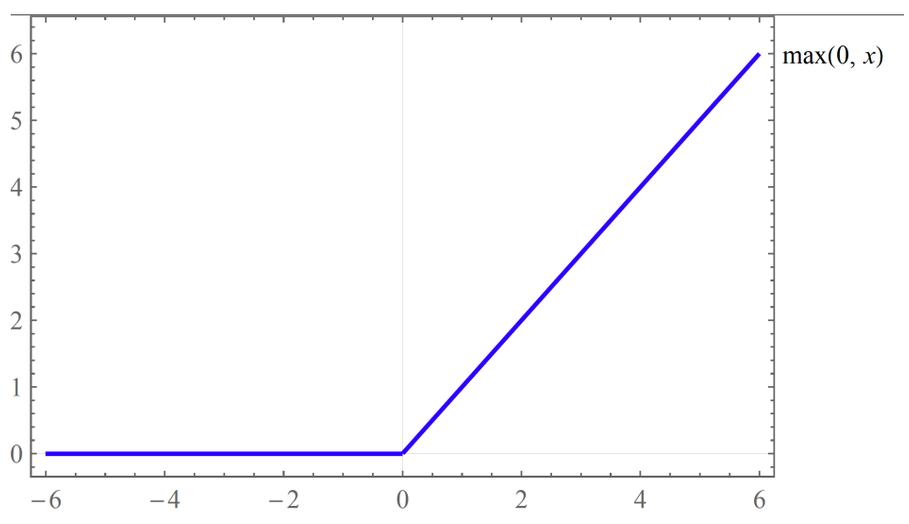
$\sigma(\langle \vec{a}, \vec{x} \rangle + b)$

Examples of activation functions





Sigmoid



ReLu

Universal approximation theorem : [Cybenko,1989]

Any continuous function can be approximated by a depth-2 neural network with sigmoid as activation function on a bounded domain

Require arbitrary width

Q: bounded width → **Depth-width trade-off**

Part I

Complexity of a function : how complicated a function is

Measures of complexity: **Oscillations** [Chatziafratis, Nagarajan, Panageas, Wang, 2020]

Fractals [Malach, Shalev-Shwartz, 2019]

Linear regions [Montufar, Pascanu, Cho, Bengio,2014]

- [There are many other reference on this topic I am less familiar with]

Topological entropy 

Definition of topological entropy [Adler, Konheim, McAndrew, 1965]

- X: a compact Hausdorff space
  f: a continuous map from X to X.
  $\mathcal{A}$: open cover of X, i.e, each element in $\mathcal{A}$ is an open subset of X and their union is X.

- $\mathcal{N}(\mathcal{A})$: minimal cardinality of the subcover from $\mathcal{A}$.

$$\mathcal{N}(\mathcal{A}) = \min\{Card(\mathcal{B}) : \mathcal{B} \subset \mathcal{A} \text{ and } \mathcal{B} \text{ is a cover of X}\},$$

where $Card(\mathcal{B})$ denotes the cardinality of $\mathcal{B}$.

- Given open covers $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n$ of X, we denote $\bigvee_{i=1}^{n} \mathcal{A}_i$ as follows,

$$\bigvee_{i=1}^{n} \mathcal{A}_i := \{A_1 \cap A_2 ... \cap A_n : A_i \in \mathcal{A}_i, \forall i, \text{ and } A_1 \cap A_2 ... \cap A_n \neq \emptyset\}.$$

Definition of topological entropy-continued

- The preimage of an open cover $\mathcal{A}$

$$\begin{aligned} f^{-i}(\mathcal{A}) &= \{f^{-i}(A) : A \in \mathcal{A}\}, \\ \mathcal{A}_f^n &= \bigvee_{i=0}^{n-1} f^{-i}(\mathcal{A}). \end{aligned}$$

Example

$\mathcal{A} = \{A_0, A_1\}$

$$\mathcal{A}_f^n = \left\{A_{i_0} \cap f^{-1}(A_{i_1}) \cap .... \cap f^{-(n-1)}(A_{i_{n-1}}) | (i_0, i_1, ..., i_{n-1}) \in \{0,1\}^n\right\}$$

Each point x in X can be encoded as $(i_0, i_1, ..., i_{n-1})$ if $x \in A_{i_0} \cap f^{-1}(A_{i_1}) \cap .... \cap f^{-(n-1)}(A_{i_{n-1}})$

$\mathcal{N}(\mathcal{A}_f^n)$ minimal number of "words" of length n needed to encode the points of X

Definition of topological entropy-continued

Def: Given a compact Hausdorff topological space X, and a continuous map $f : X \to X$, for an open cover $\mathcal{A}$, the topological entropy of f on the cover $\mathcal{A}$ is defined as

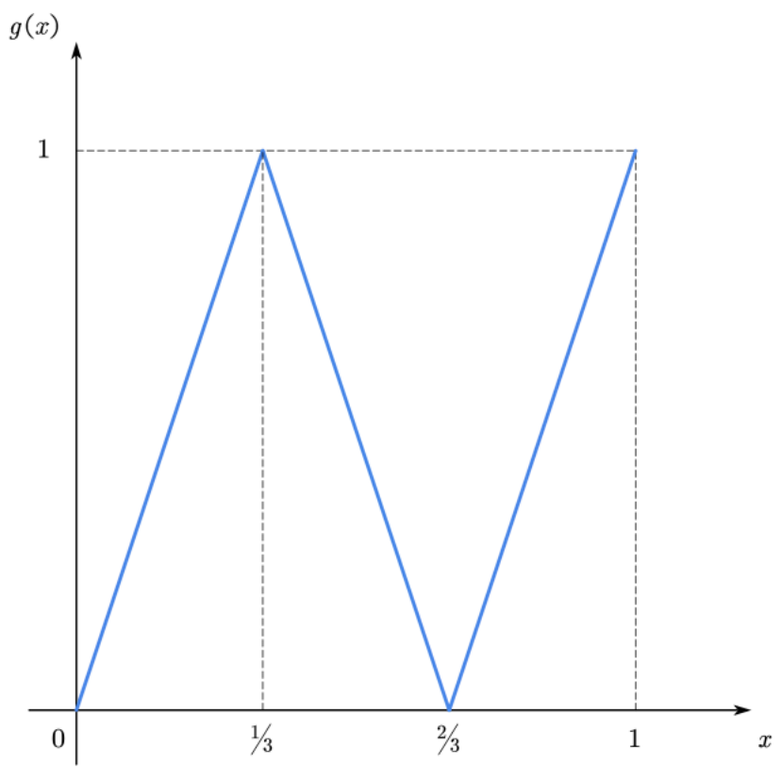
$$h_{top}(f, \mathcal{A}) = \lim_{n\to\infty} \frac{1}{n} \log_2 \mathcal{N}(\mathcal{A}_f^n).$$

The topological entropy of f is defined as

$$h_{top}(f) = \sup_{\mathcal{A}\text{: open cover of X}} h_{top}(f, \mathcal{A}). \tag{1}$$

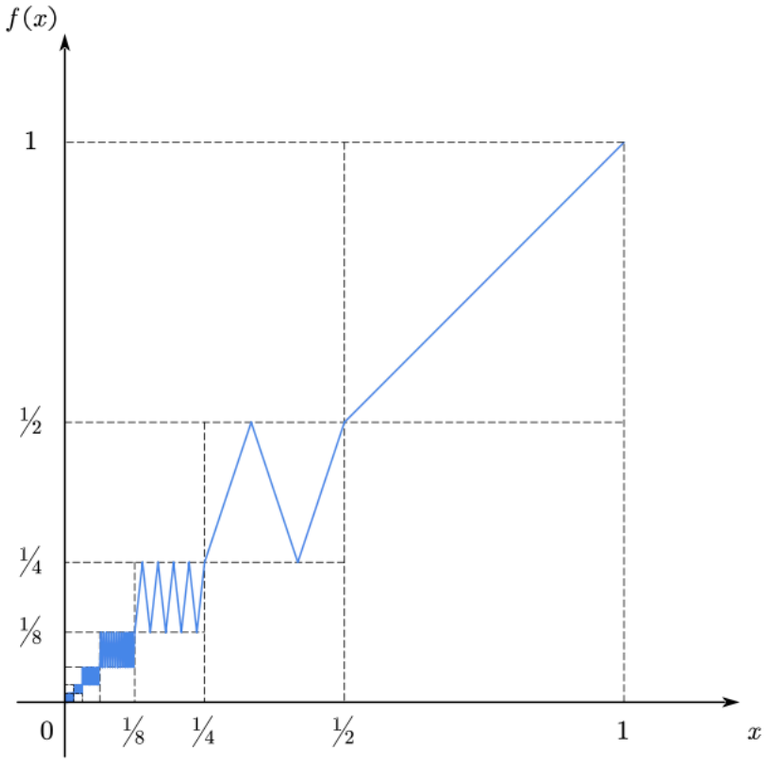
The topological entropy takes value in $[0, +\infty]$

Examples



(a)

$$h_{top}(f) = \log_2 3$$

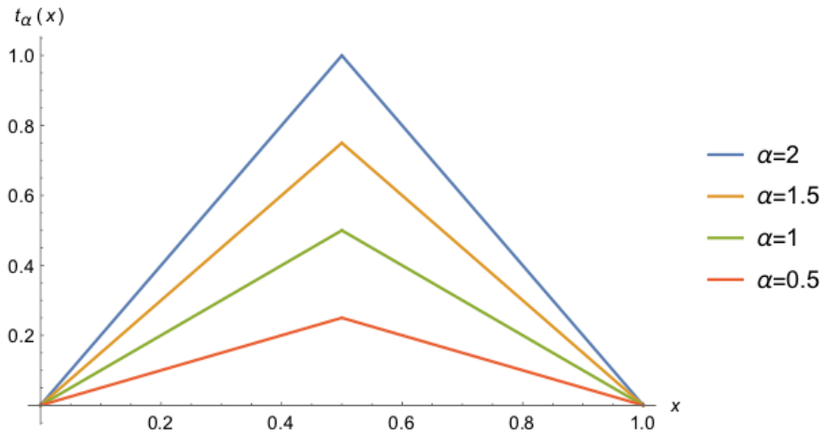


(b)

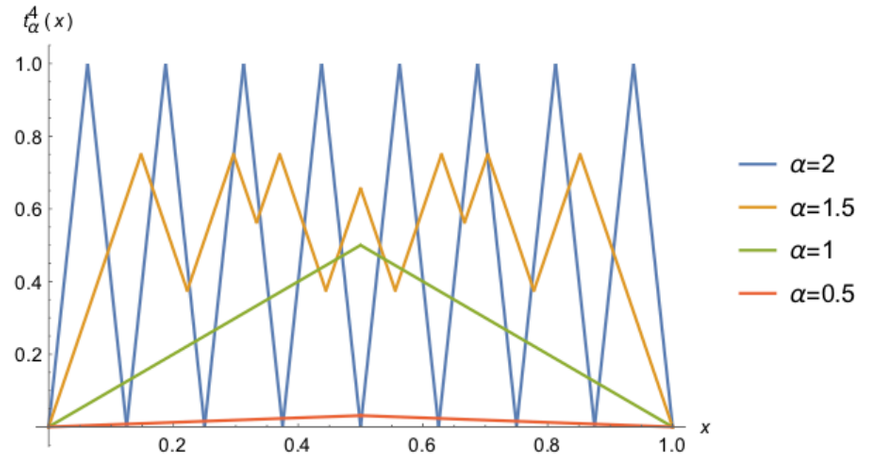
$$h_{top}(f) = +\infty$$

More Examples

$$\text{Tent map} \qquad t_\alpha(x) = \begin{cases} \alpha x, 0 \le x \le 1/2, \\ \alpha(1-x), 1/2 < x \le 1, \end{cases}$$



(a) Tent map t_α



(b) t_α^4

$$h_{top}(t_\alpha) = \begin{cases} 0, 0 \le \alpha \le 1, \\ \log_2 \alpha, 1 < \alpha \le 2. \end{cases}$$

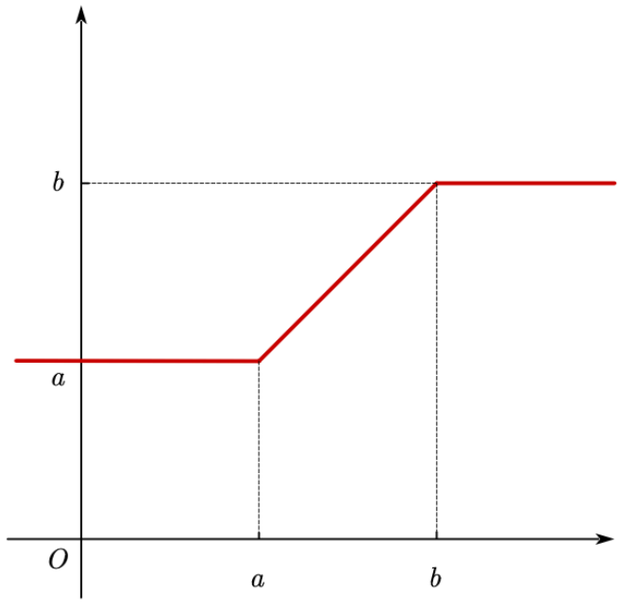
Main Result 1 (Depth-width trade-offs via topological entropy)

For any ReLU network g with at most l layers and at most m nodes per layer, then

$$h_{top}(\tau \circ g) \leq l(1 + \log_2 m).$$

Similar results also works for neural networks with other activation functions

[1] K. Bu, Y. Zhang, Q. Luo, arxiv 2010.07587



$$\tau(x) = \begin{cases} a, x \leq 1, \\ x, a \leq x \leq b \\ b, x > b. \end{cases}$$

$$\tau \circ g : [a, b] \to [a, b]$$

Part II

Complexity of a function class: richness of a function class

$$\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \}$$

$$\theta = (W_L, W_{L-1}, ..., W_1, W_0)$$

Measures of complexity of function classes

Rademacher complexity

[Bartlett and Mendelson, 2003]

VC dimension

[Vapnik and Chervonenkis, 1971]

. . .

Rademacher complexity

Def: Let $\mathcal{F}$ be a set of real-valued functions and let $S = (z_1, \ldots, z_m)$ be a set of m samples. The *(empirical) Rademacher complexity* of $\mathcal{F}$ with respect to S is

$$R_S(\mathcal{F}) = \mathop{\mathbb{E}}_{\substack{\epsilon_1,\ldots,\epsilon_m \\ \sim \mathrm{Rad}}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \left| \sum_{i=1}^{m} \epsilon_i f(z_i) \right| \right],$$

where the expectation is taken over i.i.d. Rademacher random variables, i.e., $\epsilon_i \sim \mathrm{Rad}$ for each $i \in \{1, \ldots, m\}$. Recall that the Rademacher random variable X has probability mass function

$$\Pr(X = k) = \begin{cases} 1/2 & k \in \{-1, 1\}, \\ 0 & \text{otherwise.} \end{cases}$$

[Bartlett and Mendelson, 2003]

Generalization error

Hypothesis space $\mathcal{F}$

m independent samples $\{(x_i, y_i)\}_{i=1}^m$: each (x_i, y_i) is taken i.i.d. from some unknown probability distribution D on some $\mathcal{X} \times \mathcal{Y}$

Loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

Expected error: $L(f) = \mathbb{E}_{(x,y)\sim D} l(f(x), y)$

Empirical error: $\hat{L}(f) = \frac{1}{m} \sum_{i=1}^m l(f(x_i), y_i)$

Generalization error: $|L(f) - \hat{L}(f)|$

Claim Rademacher complexity can provide an upper bound on the generalization error

[Bartlett and Mendelson, 2003]

Function class generated by quantum circuits

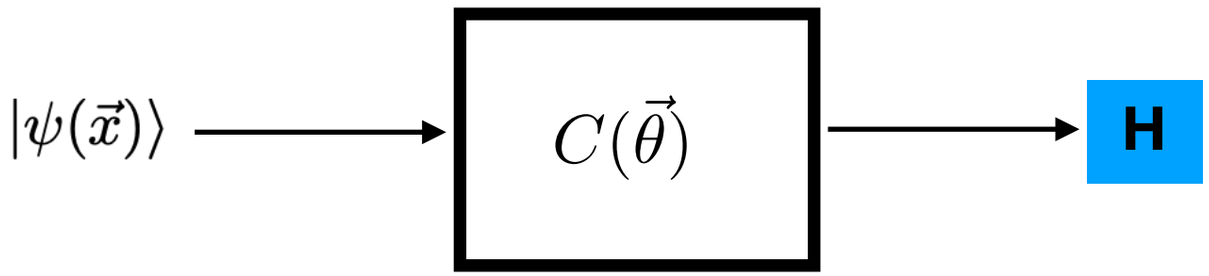
$$f_C(\vec{x}_i) = \mathrm{Tr}\left[C(|\psi(\vec{x}_i)\rangle\langle\psi(\vec{x}_i)|)H\right]$$

m independent samples $S = (\vec{x}_1, \ldots, \vec{x}_m)$ each $\vec{x}_i$ is encoded as a quantum state $|\psi(\vec{x}_i)\rangle$

C: a quantum circuit

H: an observable (Hermitian operator)

Example: $\vec{x}$ is a picture of a dog or cat, encoded as $|\psi(\vec{x})\rangle$



$f_C(\vec{x}) > 0 \rightarrow$ cat

$f_C(\vec{x}) < 0 \rightarrow$ dog

Function class generated by quantum circuits

$$f_C(\vec{x}_i) = \mathrm{Tr}\left[C(|\psi(\vec{x}_i)\rangle\langle\psi(\vec{x}_i)|)H\right]$$

$\mathcal{F} \circ \mathcal{C} := \{f_C : C \in \mathcal{C}\}$ the function class defined by the set of quantum circuits $\mathcal{C}$

Goal bound the Rademacher complexity of quantum circuit by the amount of resource (magic), depth and width of quantum circuits

Resource theory-an overview

- Free states: states that carry no resource

  Resource states: states which are not free

- Free operations: manipulations that are considered easy

  Resource operations: quantum operations which are not free

- Quantifier: function that measures how much resource in the states or operations

Resource theory of magic

Pauli operators $\mathcal{P}_n$: all tensor product of n Pauli matrices $\{I, X, Y, Z\}$ with a sign $\pm$

Stabilizer states: the simultaneous $+1$ eigenstate of 2^n Pauli operators

(Free states) ($P_i^2 = I$, $[P_i, P_j] = 0$)

e.g. $|\psi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, $\{I \otimes I, X \otimes X, Z \otimes Z, -Y \otimes Y\}$

Clifford unitary: $\{U : U\mathcal{P}_n U^\dagger \subset \mathcal{P}_n\}$.

(Free operations)

Quantify the amount of magic in quantum channel

Φ → M^{Φ} $\quad M^{\Phi}_{\vec{z}\vec{x}} = \frac{1}{2^{n_2}} \mathrm{Tr}\left[P_{\vec{z}} \Phi(P_{\vec{x}})\right]$ Representation matrix (or tranform matrix) under Pauli basis

M → $\|M\|_{p,q} = \left(\frac{1}{N_1} \sum_i \|M_i\|_p^q \right)^{1/q}$ $\quad (p,q)$ *group norm*, with $0 < p, q \leq \infty$

$$\|M_i\|_p = \left(\sum_{j=1}^{N_2} |M_{ij}|^p \right)^{1/p} \qquad M_i\text{: } i\text{-th row vector}$$

Resource measure of magic: $\left\|M^{\Phi}\right\|_{p,q}$

Proposition 12. *Given a unitary channel U, the (p,q) norm can be regarded as a resource measure satisfying the following properties*

(1) (Faithfulness) For $0 < p < 2$, we have $\left\|M^U\right\|_{p,q} \geq 1$, $\left\|M^U\right\|_{p,q} = 1$ iff U is Clifford unitary.

(1') (Faithfulness) For $p > 2$, $0 < q < \infty$, we have $\left\|M^U\right\|_{p,q} \leq 1$, $\left\|M^U\right\|_{p,q} = 1$ iff U is Cllifford unitary.

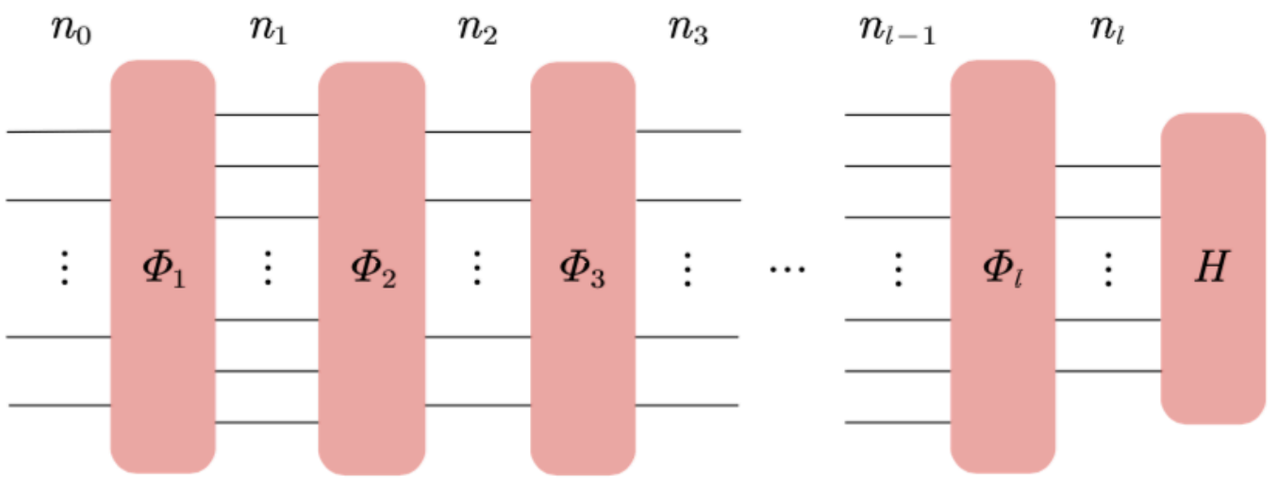
(2) (Invariance under Clifford unitaries) $\left\|M^{U_1 \circ U \circ U_2}\right\|_{p,q} = \left\|M^U\right\|_{p,q}$ for any Clifford unitaries U_1 and U_2.

(3) (Multiplicity under tensor product) $\left\|M^{U_1 \otimes U_2}\right\|_{p,q} = \left\|M^{U_1}\right\|_{p,q} \left\|M^{U_2}\right\|_{p,q}$.

(4) (Convexity) For $p \geq 1, q \geq 1$, we have $\left\|M^{\lambda U_1 + (1-\lambda) U_2}\right\|_{p,q} \leq \lambda \left\|M^{U_1}\right\|_{p,q} + (1-\lambda) \left\|M^{U_2}\right\|_{p,q}$ for $\lambda \in [0,1]$.

Resource measure for depth-l quantum circuits

Depth-l quantum circuit $\vec{C}_l := (\Phi_l, \Phi_{l-1}, \cdots, \Phi_1)$ $\quad$ i-th layer $\Phi_i : \mathcal{L}((\mathbb{C}^2)^{\otimes n_{i-1}}) \to \mathcal{L}((\mathbb{C}^2)^{\otimes n_i})$



$$\mathcal{C}^{l,\vec{n}} = \left\{ \vec{C}_l \middle| \vec{C}_l = (\Phi_l, \Phi_{l-1}, \ldots, \Phi_1), \quad \Phi_i : \mathcal{L}((\mathbb{C}^2)^{\otimes n_{i-1}}) \to \mathcal{L}((\mathbb{C}^2)^{\otimes n_i}) \right\}.$$

$\vec{n} = (n_l, \ldots, n_1, n_0)$ $\quad$ width vector

Resource measure for depth-l quantum circuits

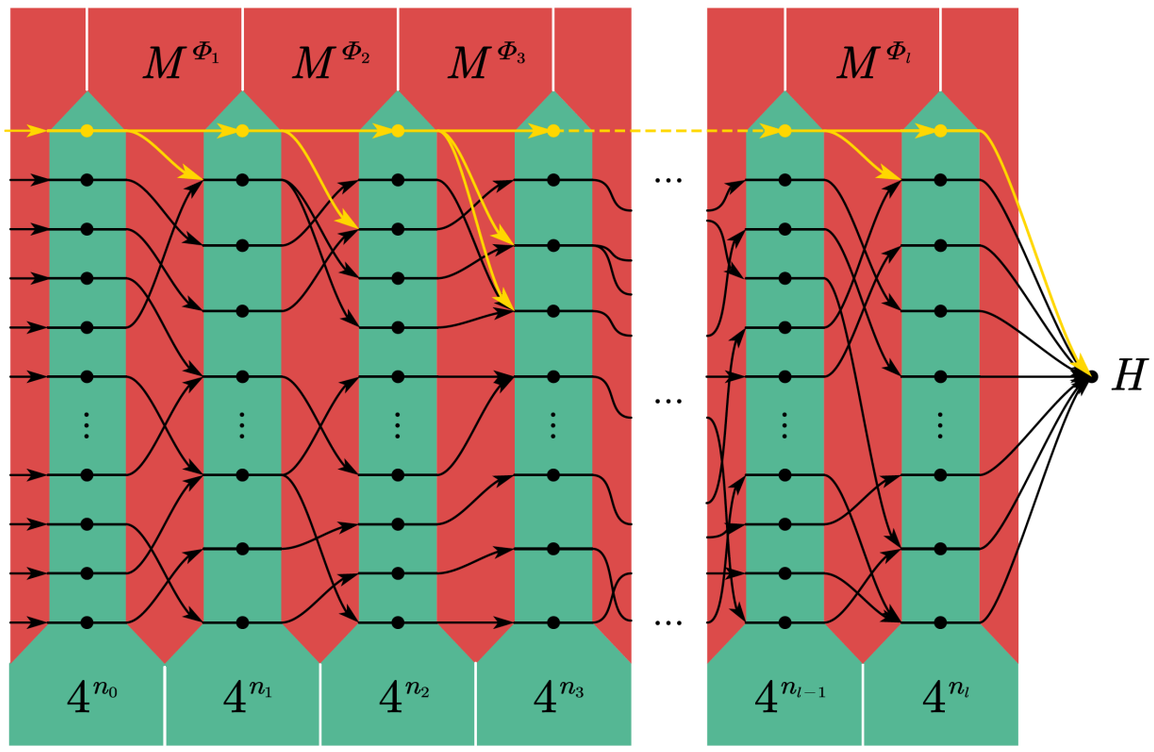
$$\nu_{p,q}(\vec{C}_l) = \frac{1}{l}\sum_{i=1}^{l} \left\| M^{\Phi_i} \right\|_{p,q}$$ average amount of magic over the layers of the quantum circuit

(Arithmetic mean)

$$\mu_{p,q}(\vec{C}_l) = \prod_{i=1}^{l} \left\| M^{\Phi_i} \right\|_{p,q},$$ (Geometric mean)

$$\mathcal{C}^{l,\vec{n}}_{\nu_{p,q}\leq\nu} := \{ \vec{C}_l \in \mathcal{C}^{l,\vec{n}} : \nu_{p,q}(\vec{C}_l) \leq \nu \},$$

$$\mathcal{C}^{l,\vec{n}}_{\mu_{p,q}\leq\mu} := \{ \vec{C}_l \in \mathcal{C}^{l,\vec{n}} : \mu_{p,q}(\vec{C}_l) \leq \mu \}$$



Main Result 2: (Upper bound on Rademacher complexity)

Rademacher complexity on m samples $S = \{\vec{x}_1, \ldots, \vec{x}_m\}$ *satisfies the following bounds*

(1) For $1 \le p \le 2$, *we have*

$$R_S(\mathcal{F} \circ \mathcal{C}^{l,\vec{n}}_{\nu_{p,q} \le \nu}) \le \nu^l 4^{\|\vec{n}\|_1 \max\{\frac{1}{p^*}, \frac{1}{q}\}} \frac{\sqrt{\min\{p^*, 8n_0\}}}{\sqrt{m}} K_p(S,H).$$

→ similar results for $\mu_{p,q}$

(2) For $2 < p < \infty$, *we have*

$$R_S(\mathcal{F} \circ \mathcal{C}^{l,\vec{n}}_{\nu_{p,q} \le \nu}) \le \nu^l 4^{\|\vec{n}\|_1 \max\{\frac{1}{p^*}, \frac{1}{q}\}} \frac{\sqrt{p^*}}{m^{1/p}} K_p(S,H)$$

[2] K. Bu, D. Koh, L. Li, Q. Luo, Y. Zhang, arxiv 2101. 06154

$$\|\vec{n}\|_1 = \sum_{i=1}^{l} n_i, \qquad \frac{1}{p} + \frac{1}{p^*} = 1$$

$$K_p(S,H) = \|\vec{\alpha}\|_p \max_i \left\| \vec{f}_I(\vec{x}_i) \right\|_{p^*}$$

$\vec{\alpha}$ *and* $\vec{f}_I(\vec{x}_i)$ *are the representation vectors of* H *and* $|\psi(x_i)\rangle\langle\psi(x_i)|$ *in the Pauli basis*

Summary

- **Topological entropy as measure of complexity of (classical) neural networks**

  → **Depth-width trade-off relationship**

- **Rademacher complexity of quantum circuits**

  → **Upper bound by the amount of magic, depth and width**

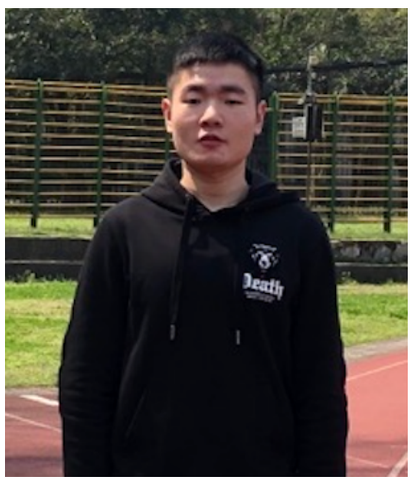
Further direction

- **Topological entropy** → **Quantum neural networks**
- **Rademacher complexity** → **Other measures**

Collaborators

Dax Enshan Koh

Lu Li



Qianxian Luo

Yaobo Zhang

Reference

[1] K. Bu, Y. Zhang, Q. Luo, arxiv 2010.07587

[2] K. Bu, D. Koh, L. Li, Q. Luo, Y. Zhang, arxiv 2101. 06154

-
- [Some other works in preparation]
-

Thank You!