Koopman-based generalization bound: New aspect for full-rank weights

Y. Hasimoto[1,2], S. Sonoda[2], I. Ishikawa[3,2], A. Nitanda[4], and T. Suzuki[5,2]

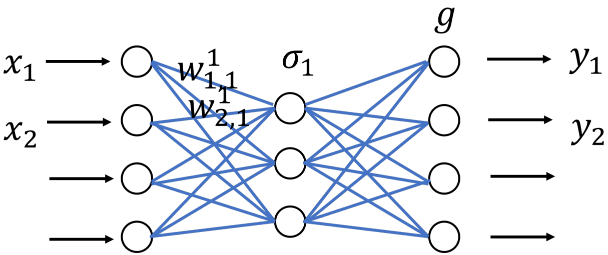
1. NTT 2. RIKEN AIP 3. Ehime University
4. A*STAR CFAR 5. The University of Tokyo

Neural Networks

Consider a neural network :

$$f = g \circ b_L \circ W_L \circ \cdots \circ \sigma_1 \circ b_1 \circ W_1 \tag{1}$$

$W_j : \mathbb{R}^{d_{j-1}} \to \mathbb{R}^{d_j}$ $(j = 1, \ldots, L)$: linear map
$b_j : \mathbb{R}^{d_j} \to \mathbb{R}^{d_j}$: bias
$\sigma_j : \mathbb{R}^{d_j} \to \mathbb{R}^{d_j}$: activation function
$g : \mathbb{R}^{d_L} \to \mathbb{C}$: nonlinear function



Representing a neural network using Koopman operators

$H_j := H^{s_j}(\mathbb{R}^{d_j})$: Sobolev space of order $s_j > d_j/2$ (RKHS)
Koopman operator K_h is the linear operator from H_j to H_{j-1} defined as

$$K_h v := v \circ h \tag{2}$$

Assumption : $g \in H_L$ and $\|K_{\sigma_j}\| < \infty$

$$f(x) = K_{W_1} K_{b_1} K_{\sigma_1} \cdots K_{W_L} K_{b_L} g(x) \tag{3}$$

We can represent the network using the product of Koopman operators.

$$\boxed{g(x)} \xrightarrow{K_{b_L}} \boxed{g(x+b_L)} \xrightarrow{K_{W_L}} \boxed{g(W_L x + b_L)}$$

$$\xrightarrow{K_{W_{L-1}} K_{b_{L-1}} K_{\sigma_{L-1}}} \boxed{g(W_L \sigma_{L-1}(W_{L-1}x + b_{L-1}) + b_L)}$$

$$\longrightarrow \cdots$$

Koopman-based generalization bound

<u>Question:</u> Do the networks with high-rank weight matrices generalize well?
$\rightarrow$ Emperically, yes! But was not fully understood theoretically.

F : Class of all functions represented by the neural network.
$\mathcal{W}_j(C,D) = \{W \in \mathbb{R}^{d_{j-1}\times d_j} \mid d_j \geq d_{j-1},\ \|W\| \leq C,\ \det(W^*W)^{\frac{1}{2}} \geq D\}$
$F_{\mathrm{inj}}(C,D) = \{f \in F \mid W_j \in \mathcal{W}_j(C,D)\}$

Theorem 1

Let $s_j > d_j/2$. The Rademacher complexity $\hat{R}_n(\mathbf{x}, F_{\mathrm{inj}}(C,D))$ is bounded as

$$\hat{R}_n(\mathbf{x}, F_{\mathrm{inj}}(C,D))) \leq O\left(\frac{\|g\|_{H_L}}{\sqrt{n}} \sup_{W_j \in \mathcal{W}_j(C,D)} \prod_{j=1}^{L} \frac{\|K_{\sigma_j}\| \|W_j\|^{s_j-1}}{\det(W_j^* W_j)^{1/4}}\right). \quad (4)$$

The factor $\|W_j\|^{s_j-1}/\det(W_j^*W_j)^{1/4}$ comes from the Koopman operator with respect to W_j.

Remarks regarding the bound

- If the weight matrices are unitary, then the bound becomes small.
- Our bound can become small even if W_j has large singular values.
- We can generalize our bound to the case where W_j is non-injective ($d_j < d_{j-1}$), but we have to modify the network slightly.
- We showed that our bound can be combined with existing bounds.
- Our bound is suitable for lower layers. We can interpret that signals are transformed on the lower layers and are extracted on the higher layers. The transformation leads to the better extraction of signals on the higher layers.