Size-Independent Sample Complexity of Neural Networks

arXiv preprint (2018)

Noah Golowich, Alexander Rakhlin and Ohad Shamir

Presenter: Gyuseung Baek

July 8, 2019

Introduction

- DNN rarely overfit training data despite of its large structure
- Calculate sample complexity(generalization error) independent to the network size is important

- Ex) Neyshabur et al. [2015]: Fully connected network, W_i: ith parameter matrix. $|||\cdot|||$: Frobenius norm, each ıth layer's Frobenius norm is bounded by $M_F(i)$, then the generalizatino error scales as

$$\mathcal{O}\left(\frac{B2^d \prod_{j=1}^{d} M_F(j)}{\sqrt{m}}\right)$$

Introduction

- Reduce complexity from exponential to polynomial depth dependence
- From depth dependence to depth independence
- Calculate lower bound

Notation

- Small letter is a vector, Capital letter is a matrix
- For $p \geq 1, \|\mathbf{w}\|_p = \left(\sum_{i=1}^{h} |\mathbf{w}_i|^p\right)^{1/p}$ will refer to the ℓ_p norm.
- For $p \geq 1, \|W\|_p$ is the Schatten p-norm (p-norm of the spectrum of W).
  - $p = \infty$: Spectral norm (we will drop the ∞ subscript).
    $p = 2$: Frobenius norm($\|W\|_F$), $p = 1$: trace norm.
- $\|W\|_{p,q} := \left(\sum_k \left(\sum_j |W_{j,k}|^p\right)^{q/p}\right)^{1/q}$

- For function class $\mathcal{H}$ and some set of data points $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{X}$, we define the (empirical) Rademacher complexity $\hat{\mathcal{R}}_m(\mathcal{H})$ as

$$\hat{\mathcal{R}}_m(\mathcal{H}) = \mathbb{E}_\epsilon \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i h(\mathbf{x}_i) \right]$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ is a vector uniformly distributed in $\{-1, +1\}^m$.

Model

- Domain: $\mathcal{X} = \{\mathrm{x} : \|\mathrm{x}\| \leq B\}$
- Model: Standard Fully connected DNN (real function)

$$\mathbf{x} \mapsto W_d \sigma_{d-1} \left(W_{d-1} \sigma_{d-2} \left(\ldots \sigma_1 \left(W_1 \mathbf{x} \right) \right) \right)$$

  where each W_j is a parameter matrix, and each σ_j is some fixed Lipschitz continuous function.
- d: depth, h: width (maximal row or column dim of $W_1, \cdots, W_d$
- $\forall j, \sigma_j$ has a Lipschitz constant of at most 1, positive-homogeneous ($\sigma(\alpha z) = \alpha \sigma(z)$ for all $\alpha \geq 0$ and $z \in \mathbb{R}$)

- W_b^r: shorthand for the matrix tuple $\{W_b, W_{b+1}, \ldots, W_r\}$
- $N_{W_b^r}$: function from layers b through r:
  $\mathrm{x} \mapsto W_r \sigma_{r-1} \left(W_{r-1} \sigma_{r-2} \left(\ldots \sigma_b \left(W_b \mathrm{x} \right) \right) \right)$

From exponential to polynomial depth dependence

- Compute the Rademacher complexity: using 'peeling' argument: reduce depth r networks to depth $r-1$ networks.

$$\mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_d} \frac{1}{m} \sum_{i=1}^{m} \epsilon_i h(\mathbf{x}_i) = \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_{d-1} W_d : \|W_d\|_F \leq M_F(d)} \frac{1}{m} \sum_{i=1}^{m} \epsilon_i W_d \sigma(h(\mathbf{x}_i))$$

  can be upper bounded by $M_F(d) \cdot \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_{d-1}} \left\| \frac{1}{m} \sum_{i=1}^{m} \epsilon_i \sigma(h(\mathrm{x}_i)) \right\| \leq 2M_F(d) \cdot \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_{d-1}} \| \frac{1}{m} \sum_{i=1}^{m} \epsilon_i h(\mathbf{x}_i)) \|$

- Factor 2 is generally unavoidable(Ledoux and Talagrand, 1991)

- By Jensen's Inequality,

$$\begin{aligned} \hat{\mathcal{R}}_m(\mathcal{H}) &= \frac{1}{\lambda} \log \exp \left(\lambda \cdot \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \epsilon_i h(\mathrm{x}_i) \right) \\ &\leq \frac{1}{\lambda} \log \left(\mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \exp \left(\lambda \sum_{i=1}^{m} \epsilon_i h(\mathrm{x}_i) \right) \right) \end{aligned}$$

From exponential to polynomial depth dependence

Theorem 1. Let $\mathcal{H}_d$ be the class of real-valued networks of depth d over the domain $\mathcal{X}$, where each parameter matrix W_j has Frobenius norm at most $M_F(j)$, and let σ be *a* 1-Lipschitz, positive-homogeneous activation function which is applied element-wise. Then

$$\hat{\mathcal{R}}_m\left(\mathcal{H}_d\right) \leq \frac{1}{m}\prod_{j=1}^{d} M_F(j) \cdot \left(\sqrt{2\log(2)d}+1\right)\sqrt{\sum_{i=1}^{m}\|\mathbf{x}_i\|^2}$$

$$\leq \frac{B(\sqrt{2\log(2)d}+1)\prod_{j=1}^{d} M_F(j)}{\sqrt{m}}$$

From exponential to polynomial depth dependence

- Thm 1 can be applied to the infinity norm

Theorem 2. Let $\mathcal{H}_d$ be the class of real-valued networks of depth d over the domain $\mathcal{X}$, where $\|W_j\|_{1,\infty} \leq M(j)$, and let σ be a 1-Lipschitz activation function with $\sigma(0) = 0$, applied element-wise. Then

$$\hat{\mathcal{R}}_m(\mathcal{H}_d) \leq \frac{1}{m} \prod_{j=1}^{d} M_F(j) \cdot (\sqrt{2 \log(2) d} + 1) \sqrt{\sum_{i=1}^{m} \|\mathbf{x}_i\|^2}$$

$$\leq \frac{B(\sqrt{2 \log(2) d} + 1) \prod_{j=1}^{d} M_F(j)}{\sqrt{m}}$$

From Depth Dependence to Independence

Theorem 3. For any $p \in [1, \infty)$, any network $N_{W_1^d}$ such that $\prod_{j=1}^{d} \|W_j\| \geq \Gamma$ and $\prod_{j=1}^{d} \|W_j\|_p \leq M$ and for any $r \in \{1, \ldots, d\}$, there exists another network $N_{\tilde{W}_1^d}$(of the same depth and layer dimensions) with the following properties:

- $\tilde{W}_1^d = \left\{\tilde{W}_1, \ldots, \tilde{W}_d\right\}$ is identical to W_1^d, except for the parameter matrix $\tilde{W}_{r'}$ in the $r' - th$ layer, for some $r' \in \{1, 2, \ldots, r\}$. The matrix $\tilde{W}_{r'}$ is of rank at most 1 , and equals $s\mathbf{u}\mathbf{v}^\top$ where $s, \mathbf{u}, \mathbf{v}$ are some leading singular value and singular vectors pairs of $W_{r'}$
- $\sup_{\mathbf{x} \in \mathcal{X}} \left\| N_{W_1^d}(\mathbf{x}) - N_{\tilde{W}_1^d}(\mathbf{x}) \right\| \leq B \left(\prod_{j=1}^{d} \|W_j\|\right) \left(\frac{2p \log(M/\Gamma)}{r}\right)^{1/p}$

(Only one parameter is different and output is similar)

From Depth Dependence to Independence

Theorem 4. Let $\mathcal{H}$ be a class of fiuctions from Euclidean space to $[-R, R]$. Let $\mathcal{F}_{L,a}$ be the class of of L -Lipschitz functions from $[-R, R]$ to $\mathbb{R}$, such that $f(0) = \text{a}$ for some fixed a. Letting $\mathcal{F}_{L,a} \circ \mathcal{H} := \{f(h(\cdot)) : f \in \mathcal{F}_{L,a}, h \in \mathcal{H}\}$, its Rademacher complexity satisfies

$$\hat{\mathcal{R}}_m\left(\mathcal{F}_{L,a} \circ \mathcal{H}\right) \leq cL\left(\frac{R}{\sqrt{m}} + \log^{3/2}(m) \cdot \hat{\mathcal{R}}_m(\mathcal{H})\right)$$

where $c > 0$ is a universal constant.

Theorem 5. Consider the following hypothesis class of networks on $\mathcal{X} = \{\mathrm{x} : \|\mathrm{x}\| \leq B\}$:

$$\mathcal{H} = \left\{ N_{W_1^d} : \begin{array}{c} \prod_{j=1}^{d} \|W_j\| \geq \Gamma \\ \forall j \in \{1 \dots d\}, W_j \in \mathcal{W}_j, \max\left\{ \frac{\|W_j\|}{M(j)}, \frac{\|W_j\|_p}{M_p(j)} \right\} \leq 1\} \end{array} \right\}$$

for some parameters $p, \Gamma \geq 1, \{M(j), M_p(j), \mathcal{W}_j\}_{j=1}^{d}$. Also, for any $r \in \{1, \dots, d\}$, define

$$\mathcal{H}_r = \left\{ N_{W_1^r} :: \begin{array}{c} N_{W_1^r} \text{ maps to } \mathbb{R} \\ \forall j \in \{1 \dots r-1\}, W_j \in \mathcal{W}_j \\ \forall j \in \{1 \dots r\}, \max\left\{ \frac{\|W_j\|}{M(j)}, \frac{\|W_j\|_p}{M_p(j)} \right\} \leq 1 \end{array} \right\}$$

From Depth Dependence to Independence

Finally, for $m > 1$, let $\ell \circ \mathcal{H} = \{(\ell_1(h(\mathrm{x}_1))) : h \in \mathcal{H}\}$, where $\ell_1, \ldots, \ell_m$ are real-valued loss functions which are $\frac{1}{\gamma}$-Lipschitz and satisfy

$\ell_1(0) = \ell = \ell_m(0) = a$, for some $a \in \mathbb{R}$. Assume that $|a| \leq \frac{B \prod_{j=1}^{d} M(j)}{\gamma}$
Then the Rademacher complexity $\hat{\mathcal{R}}_m(\ell \circ \mathcal{H})$ is upper bounded by

$$\frac{cB \prod_{j=1}^{d} M(j)}{\gamma} \min_{r \in \{1,\ldots,d\}} \left\{ \frac{\log^{3/2}(m)}{B} \cdot \max_{r' \in \{1,\ldots,r\}} \frac{\hat{\mathcal{R}}_m(\mathcal{H}_{r'})}{\prod_{j=1}^{r'} M(j)} + \left(\frac{\log\left(\frac{1}{\Gamma} \prod_{j=1}^{d} M_p(j)\right)}{r} \right)^{1/p} + \frac{1 + \sqrt{\log r}}{\sqrt{m}} \right\}$$

where $c > 0$ is a universal constant.

From Depth Dependence to Independence

Corr 1. Let $\mathcal{H}$ be the class of depth-d neural networks, where each parameter matrix W_j satisfies $\|W_j\|_F \leq M_F(j)$, and with 1-Lipschitz, positive-homogeneous, element-wise activation functions. Assuming the loss function ℓ and $\mathcal{H}$ satisfy the conditions of Thm. 5 (with the sets $\mathcal{W}_j$ being unconstrained, it holds that

$$\hat{\mathcal{R}}_m(\ell \circ \mathcal{H}) \leq \mathcal{O}\left(\frac{B\Pi_{j=1}^d M_F(j)}{\gamma} \cdot \min\left\{\frac{\log^{3/4}(m)\sqrt{\log\left(\frac{1}{\Gamma}\Pi_{j=1}^d M_F(j)\right)}}{m^{1/4}}, \sqrt{\frac{d}{m}}\right\}\right)$$

where $\log(z) := \max\{1, \log(z)\}$

From Depth Dependence to Independence

Ignoring logarithmic factors and replacing the min by its first argument, the bound in the corollary is at most

$$\tilde{\mathcal{O}}\left(\frac{B\prod_{j=1}^{d} M_F(j)}{\gamma}\sqrt{\frac{\log\left(\frac{1}{\Gamma}\prod_{j=1}^{d} M_F(j)\right)}{\sqrt{m}}}\right)$$

Assuming $\prod_j M_F(j)$ and $\prod_j M_F(j)/\Gamma$ are bounded by a constant, we get a bound which does not depend on the width or depth of the network.

Lower Bound for Schatten Norms

Thm 7. Let $\mathcal{H}$ be the class of depth-d, width-h neural networks, where each parameter matrix W_j with respect to which satisfies $\left\|W_j\right\|_p \leq M_p(j)$ for some Schatten $p-$norm $\|\cdot\|_p$ (and where use the convention that $p = \infty$ refers to the spectral norm). Then there exists a choice of $\frac{1}{\gamma}-$Lipschitz loss ℓ and data points $\mathrm{x}_1, \ldots, \mathrm{x}_m \in \mathcal{X}$ with respect to which

$$\hat{\mathcal{R}}_m(\ell \circ \mathcal{H}) \geq \Omega\left(\frac{B\prod_{j=1}^{d} M_p(j) \cdot h^{\max\left\{0, \frac{1}{2} - \frac{1}{p}\right\}}}{\gamma\sqrt{m}}\right)$$