Risk bounds for aggregated shallow neural networks using Gaussian priors

Arnak Dalalyan [2] Laura Tinsi[1,2]

[1]EDF Lab [2]CREST

Context and scope

- We aim to find theoretical guarantees that legitimate the good empirical performances of neural networks
- Most guarantees established in the literature focus on minimizing the training error but one can tackle the problem from another perspective

$\Longrightarrow$ We focus on estimators defined as a **"mixture" of a given family of weak estimators** in the PAC-Bayesian framework. We address the following questions in the setting of shallow neural networks:

1. How does the weight initialization impact the tightness of bounds ?
2. How to choose the size of the hidden layer ?
3. What kind of risk guarantee these choices induce ?

Statistical setting

- $(\mathcal{Z}, \mathscr{A})$ measurable space
- $\mathbf{Z^n} = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ realizations from an unknown distribution $\mathcal{P}$ on $(\mathcal{Z}^n, \mathscr{A}^{\otimes n})$.
- $\mathcal{X} \subset \mathbb{R}^{D_0}$, $D_0 \geq 1$, μ measure on $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$.
- $f_{\mathcal{P}} \in \mathcal{F} := \{f : \mathcal{X} \to \mathbb{R}^{D_2},\ D_2 \in \mathbb{N}\}$, function depending on $\mathcal{P}$ to estimate
- $\mathcal{F}_{\mathsf{W}} := \{f_{\boldsymbol{w}}, \boldsymbol{w} \in \mathsf{W}\} \subset \mathcal{F}$, indexed by $(\mathsf{W}, \mathscr{B}(\mathsf{W}))$, $\mathsf{W} \subset \mathbb{R}^d$.
- $\ell : \mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}_+$, standard ℓ_2 loss
- $\widehat{f}_n : \mathcal{Z}^n \mapsto \mathcal{F}_{\mathsf{W}}$ an estimator computed from the observed data $\mathbf{Z^n}$

PAC Bayesian framework

This theory originates from:

$\hookrightarrow$ **Probably Approximately Correct bounds** that are bounds in probability

$\hookrightarrow$ **Generalized Bayesian learning** that for a prior distribution π over W, defines a posterior distribution $\widehat{\pi}_n(\boldsymbol{w}|\mathbf{Z^n}) \propto \mathscr{L}_{\boldsymbol{w},n}(\mathbf{Z^n})\pi(\boldsymbol{w})$ where a given loss functional $\mathscr{L}_{\boldsymbol{w},n}$, measures the performance of a function $f_{\boldsymbol{w}}$ given $\mathbf{Z^n}$

We can then define the mean aggregate estimator:

$$\widehat{f}_n = \int_{\mathsf{W}} f_{\boldsymbol{w}} \widehat{\pi}_n(d\boldsymbol{w}). \tag{1}$$

Bounds in expectation

Under some assumptions, for $\widehat{f}_n$ defined as in (1), the following inequality holds

$$\mathbf{E}_{\mathcal{P}}[\|\widehat{f}_n - f_{\mathcal{P}}\|^2_{\mathbb{L}_2}] \leq C \inf_{p \in \mathcal{P}_{\mathsf{W}}} \left\{ \int_{\mathsf{W}} \|f_{\boldsymbol{w}} - f_{\mathcal{P}}\|^2_{\mathbb{L}_2(\mu)}\, p(d\boldsymbol{w}) + \frac{\beta}{n} D_{\mathrm{KL}}(p\|\pi) \right\}. \tag{2}$$

where π is the prior, β a temperature parameter, C a universal constant and $\mathcal{P}_{\mathsf{W}}$ the set of distributions over W.

Aggregated shallow neural networks

- Neural networks with one hidden layer are a particular specification of the subset $\mathcal{F}_{\mathsf{W}}$, where W defines the weights of the neural network
- W can then be divided into the weights $\boldsymbol{w_1}$ of the hidden layer , and the weights $\boldsymbol{w_2}$ of the output layer, so that $\boldsymbol{w_1} \in \mathbb{R}^{D_0 \times D_1}$, $\boldsymbol{w_2} \in \mathbb{R}^{D_1 \times D_2}$, and the overall dimension is $d = D_1(D_0 + D_2)$

The neural network parametrized by $\boldsymbol{w}$ has the form:

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}_2^\top \bar{\sigma}(\boldsymbol{w}_1^\top \boldsymbol{x}) \in \mathbb{R}^{D_2}, \quad \forall \boldsymbol{x} \in \mathbb{R}^{D_0} \quad \text{with } \bar{\sigma} : \boldsymbol{x} \in \mathbb{R}^{D_1} \mapsto \begin{bmatrix} \sigma(x_1) \\ \vdots \\ \sigma(x_{D_1}) \end{bmatrix} \in \mathbb{R}^{D_1}, \tag{3}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function.

Assumption $(\sigma{-}L)$ there exists $L_\sigma > 0 |\ \forall x, y \in \mathbb{R}, |\sigma(x) - \sigma(y)| \leq L_\sigma |x - y|$.

Step 1: Oracle bound for a Gaussian prior

Other formulation of the PAC Bayesian inequality

Let $\boldsymbol{w}^* \in \mathrm{argmin}_{\boldsymbol{w} \in \mathsf{W}} \|f_{\boldsymbol{w}} - f_{\mathcal{P}}\|_{\mathbb{L}_2(\mu)}$, using the triangle inequalities, (2) yields:

$$\boxed{\left(C^{-1}\, \mathbf{E}_{\mathcal{P}}\left[\|\widehat{f}_n - f_{\mathcal{P}}\|^2_{\mathbb{L}_2(\mu)}\right]\right)^{1/2} \leq \underbrace{\|f_{\boldsymbol{w}^*} - f_{\mathcal{P}}\|_{\mathbb{L}_2(\mu)}}_{\text{approximation error}} + \underbrace{\mathrm{Rem}_n(\boldsymbol{w}^*)^{1/2}}_{\text{estimation error}}} \tag{4}$$

with the remainder term given by

$$\mathrm{Rem}_n(\bar{\boldsymbol{w}}) \triangleq \inf_{p \in \mathcal{P}_{\mathsf{W}}} \left\{ \int_{\mathsf{W}} \|f_{\boldsymbol{w}} - f_{\bar{\boldsymbol{w}}}\|^2_{\mathbb{L}_2(\mu)}\, p(d\boldsymbol{w}) + \frac{\beta}{n} D_{\mathrm{KL}}(p\|\pi) \right\}. \tag{5}$$

$\Longrightarrow$ **Main goal = analyze the estimation error.** For this, we proceed in 3 steps:

1. We assume the prior distribution and the set $\mathcal{P}_{\mathsf{W}}$ are spherical Gaussian distributions
2. Replace the infinimum in (5) by a suitably chosen p
3. **Tune the variance of the prior** π, so that the worst-case value of the remainder, $\sup_{\bar{\boldsymbol{w}} : \|\bar{\boldsymbol{w}}_\ell\|_{\mathsf{F}} \leq B_\ell} \mathrm{Rem}_n(\bar{\boldsymbol{w}})$ is minimized.

Oracle inequality

For a method of aggregation of shallow neural networks $\widehat{f}_n$ as (1) we may obtain under some assumptions:

$$\boxed{\left(C^{-1}\mathbf{E}_{\mathcal{P}}\left[\|\widehat{f}_n - f_{\mathcal{P}}\|^2_{\mathbb{L}_2(\mu)}\right]\right)^{1/2} \leq \|f_{\boldsymbol{w}*} - f_{\mathcal{P}}\|_{\mathbb{L}_2(\mu)} + \left\{\frac{\beta d}{n}\tilde{g}(n/d)\right\}^{1/2}} \tag{6}$$

where $\tilde{g}$ is at most of logarithmic growth.

Step 2: Tuning of the hidden layer

Sigmoid activation functions

Maiorov and Meir (2000) provide approximation results over Sobolev smoothness classes $W_2^r([0,1]^{D_0})$ for $D_2 = 1$ such that we can rewrite (6):

$$\mathbf{E}_{\mathcal{P}}\left[\|\widehat{f}_n - f_{\mathcal{P}}\|^2_{\mathbb{L}_2(\mu)}\right] \leq g(D_1) D_1^{-2r/D_0} + \bar{g}(n/d)\frac{D_1 D_0}{n},$$

where $g,\ \bar{g}$ are at most logarithmic functions.

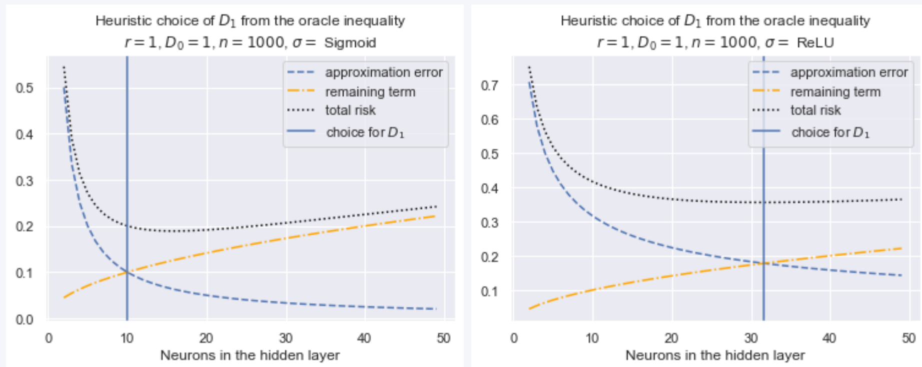
Relu activation functions

Siegel and Xu (2020) provide approximation results over Sobolev smoothness classes $W_2^r([0,1]^{D_0})$ for $D_2 = 1$ such that we can rewrite (6):

$$\mathbf{E}_{\mathcal{P}}\left[\|\widehat{f}_n - f_{\mathcal{P}}\|^2_{\mathbb{L}_2(\mu)}\right] \leq g(D_1) D_1^{-2\bar{r}/(D_0+1)} + \bar{g}(n/d)\frac{D_1 D_0}{n},$$

for $r > \bar{r} \geq \frac{D_0}{2}$ and where $g,\ \bar{g}$ are at most logarithmic functions.

Figure 1:Approximation/estimation error trade-off for sigmoid and ReLU activation functions



$\Longrightarrow$ Good choices of D_1 lead to the following orders for the risk bounds

Result: worst-case risk bounds

Sigmoid activation functions

- *Risk bound of order $O(n^{-2r/2r+D_0})$: we reach the optimal minimax rate*
- Improves existing results on shallow neural networks and competes with deep networks

Relu activation functions

- *Risk bound of order $O(n^{-2\bar{r}/(2\bar{r}+D_0+1)})$*
- Slightly worse than the optimal minimax rate proved for deep networks but improves existing results for shallow neural networks

References

VE Maiorov and Ron Meir. On the near optimality of the stochastic approximation of smooth functions by neural networks. *Advances in Computational Mathematics*, 13(1):79--103, 2000.

Jonathan W Siegel and Jinchao Xu. High-order approximation rates for neural networks with relu ^{k} activation functions. *arXiv preprint arXiv:2012.07205*, 2020.

Arnak S Dalalyan and Alexandre B Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *International Conference on Computational Learning Theory*, pages 97--111. Springer, 2007.

G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396--3410, 2006.