Nonasymptotic theory for two-layer neural networks: Beyond the bias–variance trade-off

Huiyuan Wang and Wei Lin*

School of Mathematical Sciences, Peking University

huiyuan.wang@pku.edu.cn and weilin@math.pku.edu.cn

Abstract

Large neural networks have proved remarkably effective in modern deep learning practice, even in the overparametrized regime where the number of active parameters is much larger than the sample size. This contradicts the classical perspective that a machine learning model must trade off bias and variance for optimal generalization. To resolve this conflict, we present a nonasymptotic generalization theory for two-layer neural networks with ReLU activation function by incorporating scaled variation regularization. Interestingly, the regularizer is equivalent to ridge regression from the angle of gradient-based optimization, but plays a similar role to the group lasso in controlling the model complexity. By exploiting this "ridge–lasso duality," we obtain new prediction bounds for all network widths, which reproduce the double descent phenomenon. Moreover, the overparametrized minimum risk is lower than the underparametrized minimum risk when the signal is strong, and nearly attains the minimax optimal rate over a suitable class of functions. By contrast, we show that overparametrized random feature models suffer from the curse of dimensionality and thus are suboptimal.

1. Main Results

(i) *Functional analysis of learning limit for two-layer ReLU networks.*

- ◇ We the space consisting of infinite-width two-layer ReLU networks

$$\mathcal{G}=\left\{\mathbf{x}\mapsto\int_{\mathbb{R}^{d+1}}\big(\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b)\big)\,d\alpha(\mathbf{w}):\int_{\mathbb{R}^{d+1}}\|\mathbf{v}\|_2\,d|\alpha|(\mathbf{w})<\infty\right\}\qquad(1)$$

  and equipped it with a norm $\|f\|_{\mathcal{S}}=\int_{\mathbb{R}^{d+1}}\|\mathbf{v}\|_2\,d|\alpha_f|(\mathbf{w})$, where the signed measure $\alpha_f\in\mathcal{M}_2(\mathbb{R}^{d+1})$ is uniquely determined by

$$f(\mathbf{x})=\int_{\mathbb{R}^{d+1}}\big(\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b)\big)\,d\alpha_f(\mathbf{w})+f(\mathbf{0}).$$

- ◇ Following [1], we proved that functions in $\mathcal{G}$ are learning limits of two-layer ReLU networks; that is, $\overline{R}(f)<\infty$ if and only if $\|f\|_{\mathcal{S}}<\infty$, where

$$\overline{R}(f)=\lim_{\varepsilon\to0}\left(\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}C(\boldsymbol{\theta})\ \text{ s.t. }\ \sup_{\|\mathbf{x}\|\le\varepsilon^{-1}}|g(\mathbf{x};\boldsymbol{\theta})-f(\mathbf{x})|\le\varepsilon,\ \text{ and }\ g(\mathbf{0};\boldsymbol{\theta})=f(\mathbf{0})\right),$$

  $g(\mathbf{x};\boldsymbol{\theta})=\sum_{k=1}^m a_k\sigma(\mathbf{v}_k^T\mathbf{x}+b_k)$ denotes a finite-width ReLU network and $C(\boldsymbol{\theta})=\sum_{k=1}^m\big(\|\mathbf{v}_k\|_2^2+|a_k|^2\big)$ is the sum of squared norm of parameters.
- ◇ For a finite-width network $g(\cdot;\boldsymbol{\theta})$, we define its *scaled variation norm* as $\nu(\boldsymbol{\theta})=\sum_{k=1}^m|a_k|\|\mathbf{w}_k\|_2$ where $\mathbf{w}=(\mathbf{v}_k^T,b_k)^T$, $\forall k$.

(ii) *Prediction bounds of two-layer ReLU networks for arbitrary width.* In order to learn f^* from the training sample, we adopt the penalized empirical risk minimization (ERM) framework and seek

$$\widehat{\boldsymbol{\theta}}\in\mathop{\arg\min}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_m}J_n(\boldsymbol{\theta};\lambda)=\frac{1}{2n}\sum_{i=1}^n\big(y_i-g(\mathbf{x}_i;\boldsymbol{\theta})\big)^2+\lambda\nu(\boldsymbol{\theta}),\qquad(2)$$

where $\lambda>0$ is a regularization parameter. Under Conditions (C1)–(C3), the regularized network estimator $g(\cdot;\widehat{\boldsymbol{\theta}})$ with λ being optimally tuned, satisfies

$$\|g(\cdot;\widehat{\boldsymbol{\theta}})-f^*\|_2^2\le C\left[\|f^*\|_{\mathcal{S}}^2m^{-(d+3)/d}+(\sigma_\varepsilon^2+\|f^*\|_{\mathcal{S}}^2)\min\left\{\sqrt{\frac{d\log(en/d)}{n}},\frac{md\log(en/d)}{n}\right\}\right]\qquad(3)$$

with probability at least $1-O(n^{-C_1})$ for some constants $C_1,C>0$.

(iii) *Reproduce the double descent phenomenon.*

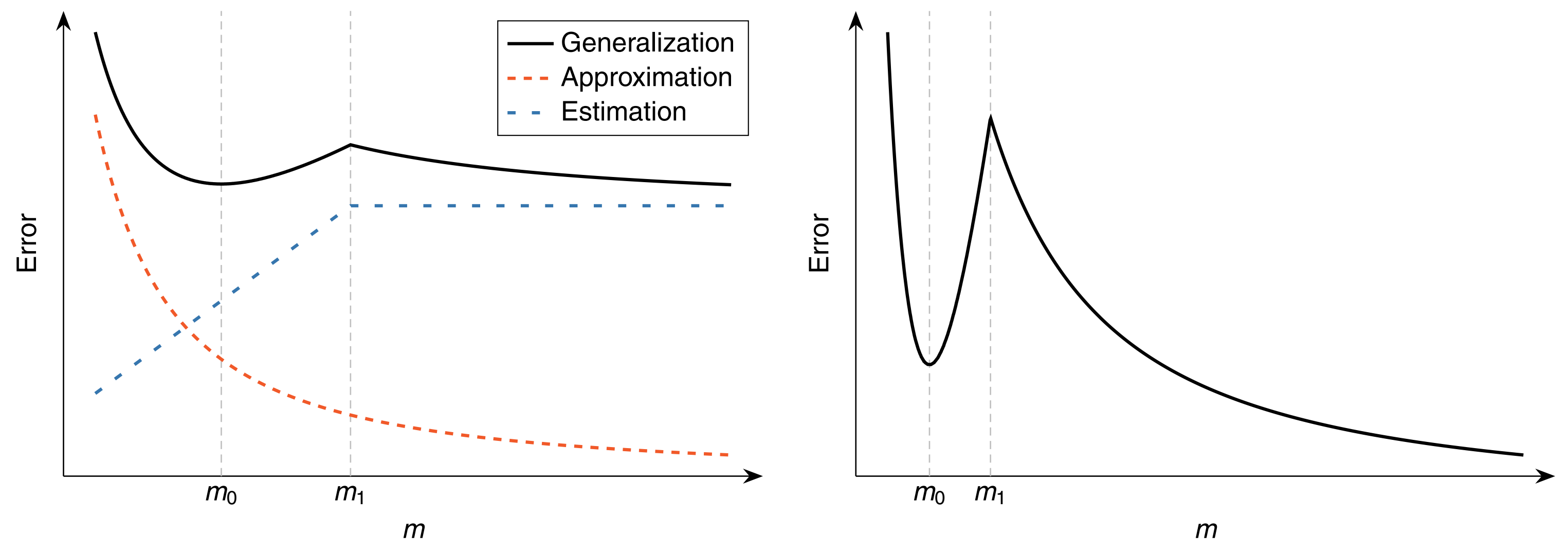

Figure 1: *Risk curves for varying network width m from the prediction bound (3) with $\|f^*\|_{\mathcal{S}}^2/\sigma_\varepsilon^2=1$, $d=6$, and $n=1000$. The left panel shows the decomposition of prediction error into approximation and estimation errors. The right panel shows the same plot but with larger m, from which it is apparent that the second valley is lower than the first.*

- ◇ The first valley or underparametrized minimum risk is $O((d\log n/n)^{(d+3)/(2d+3)})$, which occurs at $m_0\asymp(n/(d\log n))^{d/(2d+3)}$.
- ◇ The second valley or *overparametrized minimum risk* is $O(\sqrt{d\log n/n})$, which is *slightly larger* than the *underparametrized minimum risk*.
- ◇ In finite samples, however, this comparison can be reversed. The second valley is *lower than* the first whenever

$$\kappa\equiv\frac{\|f^*\|_{\mathcal{S}}^2}{\sigma_\varepsilon^2+\|f^*\|_{\mathcal{S}}^2}>\left(\frac{1}{2}\right)^{(2d+3)/d}\left(\frac{n}{d\log n}\right)^{3/(2d)}.\qquad(4)$$

  When $d\gg\log n$, the above condition approximately becomes $\kappa>1/4$, or *the signal-to-noise ratio* $\|f^*\|_{\mathcal{S}}^2/\sigma_\varepsilon^2=\kappa/(1-\kappa)>1/3$.

(iv) *Minimax optimality analyses.*

- ◇ Underparametrized neural networks attain the minimax optimal lower bound over $\mathcal{G}_M=\{f\in\mathcal{G}\colon\|f\|_{\mathcal{S}}\le M\}$ which is proved in [2] for a fixed $M>0$.
- ◇ Overparametrized neural networks attains the minimax optimal lower bound over $\cup_{M\ge0}\mathcal{G}_M$: Under some assumptions on distributions of design and noises, there exists a constant $C>0$ such that

$$\inf_{\widehat{f}}\sup_{f^*\in\mathcal{G}}\mathbb{E}\|\widehat{f}-f^*\|_2^2\ge\frac{C}{\sqrt{n\log n}},$$

  where the infimum is taken over all estimators.

In contrast to two-layer neural networks, random feature models with *optimally tuned* hyper-parameter *suffer from the curse of dimensionality*. Minimizing the ℓ_2-regularized empirical risk

$$\frac{1}{2n}\sum_{k=1}^m\big(y_i-h_{\rho_0}(\mathbf{x}_i;\mathbf{a})\big)^2+\frac{\lambda}{2}\|\mathbf{a}\|_2^2$$

gives the ridge estimator $h_{\rho_0}(\cdot;\widehat{\mathbf{a}}(\lambda))$, where $h_{\rho_0}(\mathbf{x};\mathbf{a})=\frac{1}{\sqrt{m}}\sum_{k=1}^m a_k\sigma(\mathbf{v}_k^T\mathbf{x}+b_k)$ denotes the random feature model. Under Conditions (C2) and (C3), there exists a universal constant $C>0$ such that

$$\sup_{f^*\in\mathcal{G}_M}\inf_{\lambda>0}\mathbb{E}\|h_{\rho_0}(\cdot;\widehat{\mathbf{a}}(\lambda))-f^*\|_2^2\ge\frac{CM}{d(\min(m,n))^{1/d}}.$$

2. Assumptions

The samples are independently drawn from the model

$$y_i=f^*(\mathbf{x}_i)+\varepsilon_i,\quad i=1,\dots,n,$$

subject to the following conditions:

(C1) $f^*\in\mathcal{G}_M\equiv\{f\in\mathcal{G}:\|f\|_{\mathcal{S}}\le M\}$ for some constant $M>0$;
(C2) $\mathbf{x}_i\sim\mu$ independently, where μ is supported in $\mathbb{B}^d=\{\mathbf{x}\in\mathbb{R}^d\colon\|\mathbf{x}\|_2\le1\}$;
(C3) $\varepsilon_i\sim N(0,\sigma_\varepsilon^2)$ independently and are independent of $\mathbf{x}_i$.

3. Ridge–Lasso duality of the scaled variation regularization

(i) *Equivalence to ridge regression.* The optimization problem (2) is equivalent to

$$\widehat{\boldsymbol{\theta}}_{\ell_2}=\mathop{\arg\min}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_m}\left\{\frac{1}{2n}\sum_{i=1}^n\big(y_i-g(\mathbf{x}_i;\boldsymbol{\theta})\big)^2+\frac{\lambda}{2}\sum_{k=1}^m(a_k^2+\|\mathbf{w}_k\|_2^2)\right\}.\qquad(5)$$

Consider the reparametrization $\widetilde{\boldsymbol{\theta}}=\mathcal{T}_1(\boldsymbol{\theta})$ defined by $\widetilde{a}_k=a_k\sqrt{\frac{\|\mathbf{w}_k\|_2}{|a_k|}}$, $\widetilde{\mathbf{w}}_k=\mathbf{w}_k\sqrt{\frac{|a_k|}{\|\mathbf{w}_k\|_2}}$ if $|a_k|\|\mathbf{w}_k\|_2\ne0$, and $(\widetilde{a}_k,\widetilde{\mathbf{w}}_k^T)=\mathbf{0}$ otherwise. After the reparametrization, we have $|\widetilde{a}_k|=\|\widetilde{\mathbf{w}}_k\|_2$ and the regularizer becomes $\sum_{k=1}^m|\widetilde{a}_k|\|\widetilde{\mathbf{w}}_k\|_2=\frac{1}{2}\sum_{k=1}^m(\widetilde{a}_k^2+\|\widetilde{\mathbf{w}}_k\|_2^2)$. Meanwhile, the positive homogeneity of the ReLU function implies that $a_k\sigma((\mathbf{x}_k^T,1)\mathbf{w}_k)=\widetilde{a}_k\sigma((\mathbf{x}_k^T,1)\widetilde{\mathbf{w}}_k)$, so that the network function is invariant under the reparametrization. This observation can be characterized by the following propositions.

- ◇ *Any* solution $\widehat{\boldsymbol{\theta}}_{\ell_2}$ to the optimization problem (5) is a solution to the problem (2). Conversely, if $\widehat{\boldsymbol{\theta}}$ is a solution to the optimization problem (2), then $\mathcal{T}_1(\widehat{\boldsymbol{\theta}})$ is a solution to the problem (5).
- ◇ Consider the gradient flow for the optimization problem (2) defined by $\frac{d}{dt}\boldsymbol{\theta}(t)=-\nabla_{\boldsymbol{\theta}}J_n(\boldsymbol{\theta}(t);\lambda)$ and for the problem (5) defined similarly, both initialized at $\boldsymbol{\theta}(0)=\mathcal{T}_1(\boldsymbol{\theta}_0)$ for an arbitrary $\boldsymbol{\theta}_0\in\boldsymbol{\Theta}_m$. Then the trajectories of the two gradient flows *coincide*.

(ii) *Connection to the group lasso.* Denote by $\mathbf{X}=((\mathbf{x}_1^T,1)^T,\dots,(\mathbf{x}_n^T,1)^T)^T$ the $n\times(d+1)$ design matrix, and $\mathbf{D}=\mathrm{diag}(I(\mathbf{Xw}\ge0))$ the diagonal indicator matrix for the positivity of $\mathbf{Xw}$. Consider the hyperplanes in $\mathbb{R}^{d+1}$ passing through the origin and orthogonal to $\mathbf{x}_i$, defined by $\mathbf{x}_i^T\mathbf{v}+b=0$. These n hyperplanes divide the parameter space $\mathbb{R}^{d+1}$ into finitely many regions, denoted by $R_1,\dots,R_p$, such that $\mathbf{D}$ stays constant over (the interior of) each R_j. The number of these regions, p, is at most $2\sum_{j=0}^d\binom{n-1}{j}=O\Big(d\big(\frac{en}{d}\big)^d\Big)$. Taking into account the sign of a, we thus partition the parameter space $\mathbb{R}^{d+2}$ for $(a,\mathbf{w}^T)^T$ into $2p$ regions

$$Q_j=[0,\infty)\times R_j,\quad Q_{p+j}=(-\infty,0)\times R_j,\quad j=1,\dots,p,$$

and define $\mathbf{D}_{p+j}=-\mathbf{D}_j$. Clearly, R_j and Q_j are convex cones. The linearity of the ReLU function over each Q_j and the optimality of $\widehat{\boldsymbol{\theta}}$ entail the following collinearity property.

- ◇ For any solution $\widehat{\boldsymbol{\theta}}=(\widehat{a}_1,\dots,\widehat{a}_m,\widehat{\mathbf{w}}_1^T,\dots,\widehat{\mathbf{w}}_m^T)^T$ to the optimization problem (2), if $(\widehat{a}_k,\widehat{\mathbf{w}}_k^T)^T$ and $(\widehat{a}_\ell,\widehat{\mathbf{w}}_\ell^T)^T$ lie in the same cone Q_j, then $\widehat{\mathbf{w}}_k$ and $\widehat{\mathbf{w}}_\ell$ must be *collinear*, that is, $\widehat{\mathbf{w}}_k=c_0\widehat{\mathbf{w}}_\ell$ for some constant $c_0>0$.
- ◇ Consider the "conewise collinearization" reparametrization $\widetilde{\boldsymbol{\theta}}=\mathcal{T}_2(\boldsymbol{\theta})$ defined by $\widetilde{a}_k=s_j$, $\widetilde{\mathbf{w}}_k=\frac{1}{|S_j|}\sum_{\ell\in S_j}|a_\ell|\mathbf{w}_\ell$, $k\in S_j$, where $S_j=\{1\le k\le m:(a_k,\mathbf{w}_k^T)^T\in Q_j\}$, $s_j=1$, and $s_{p+j}=-1$ for $j=1,\dots,p$. . Collect the network weights falling within the same cone and define the aggregated parameters $\mathbf{B}(\boldsymbol{\theta})=(\boldsymbol{\beta}_1(\boldsymbol{\theta}),\dots,\boldsymbol{\beta}_{2p}(\boldsymbol{\theta}))$ with $\boldsymbol{\beta}_j(\boldsymbol{\theta})=\sum_{k\in S_j}|a_k|\mathbf{w}_k$. For any $\boldsymbol{\theta}\in\boldsymbol{\Theta}_m$, the reparametrization $\mathcal{T}_2$ in satisfies $g(\mathbf{x}_i;\boldsymbol{\theta})=g(\mathbf{x}_i;\mathcal{T}_2(\boldsymbol{\theta}))$ for $i=1,\dots,n$ and $\|\mathbf{B}(\mathcal{T}_2(\boldsymbol{\theta}))\|_{2,1}=\nu(\mathcal{T}_2(\boldsymbol{\theta}))\le\nu(\boldsymbol{\theta})$. Moreover, the *solution $\widehat{\boldsymbol{\theta}}$ to the optimization problem* (2) satisfies

$$J_n(\widehat{\boldsymbol{\theta}};\lambda)=\frac{1}{2n}\left\|\mathbf{y}-\sum_{j=1}^{2p}\mathbf{D}_j\mathbf{X}\boldsymbol{\beta}_j(\widehat{\boldsymbol{\theta}})\right\|_2^2+\lambda\|\mathbf{B}(\widehat{\boldsymbol{\theta}})\|_{2,1}.$$

4. Nonasymptotic generalization bounds

(i) *Underparametrized regime.* Under Conditions (1)–(3), if $m<n/(d\log(en/d))$, then the regularized network estimator $g(\cdot;\widehat{\boldsymbol{\theta}})$ with $\lambda=\lambda_2\equiv C_1\sigma_\varepsilon\max\{m^{-(d+3)/d},md\log(en/d)/n\}$ satisfies

$$\|g(\cdot;\widehat{\boldsymbol{\theta}})-f^*\|_2^2\le C\left\{\|f^*\|_{\mathcal{S}}^2m^{-(d+3)/d}+(\sigma_\varepsilon^2+\|f^*\|_{\mathcal{S}}^2)\frac{md\log(en/d)}{n}\right\}$$

with probability at least $1-O(n^{-C_2})$ for some constants $C_1,C_2,C>0$.

(ii) *Overparametrized regime.* Under Conditions (1)–(3), if $m\ge C_1(n\log n/d)^{d/(2(d+3))}$, then the regularized network estimator $g(\cdot;\widehat{\boldsymbol{\theta}})$ with $\lambda=\lambda_1\equiv C_2\sigma_\varepsilon\sqrt{d\log(en/d)/n}$ satisfies

$$\|g(\cdot;\widehat{\boldsymbol{\theta}})-f^*\|_2^2\le C\left\{\|f^*\|_{\mathcal{S}}^2m^{-(d+3)/d}+(\sigma_\varepsilon^2+\|f^*\|_{\mathcal{S}}^2)\sqrt{\frac{d\log(en/d)}{n}}\right\}$$

with probability at least $1-O(n^{-C_3})$ for some constants $C_1,C_2,C_3,C>0$.

References

[1] Ongie, G., Willett, R., Soudry, D. and Srebro, N. (2020) A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *International Conference on Learning Representations*.

[2] Parhi, R. and Nowak, R. D. (2022). Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Trans. Inf. Theory*.

北京大学研究生科研成果交流展示会