Appendix

A Proof of Proposition 1

We first introduce some useful notations. Recall the random matrix $\mathbf{A} \in \mathbb{R}^{KM \times N}$ defined entry-wise by $\mathbf{A}_{j,k} = \phi(\boldsymbol{\omega}_k, \mathbf{x}_j)$ for all $j \in [KM]$ and $k \in [N]$ and $\mathbf{W} \in \mathbb{R}^{N \times N}$ be a diagonal matrix with entries $\mathbf{W}_{j,j} = \|\boldsymbol{\omega}_j\|$. Recall the observation vector $\mathbf{y} = [Y_1, \dots, Y_{KM}] \in \mathbb{R}^{KM}$. We introduce a new matrix $\mathbf{A}_\lambda \in \mathbb{R}^{(KM+N)\times N}$ and a new observation vector $\widetilde{\mathbf{y}} \in \mathbb{R}^{KM+N}$ as follows:

$$\mathbf{A}_\lambda = \begin{bmatrix} \mathbf{A} \\ \sqrt{\lambda KM}\mathbf{W} \end{bmatrix} \in \mathbb{R}^{(KM+N)\times N}, \quad \text{and } \widetilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \in \mathbb{R}^{kM+N}$$

Then the solution vector $\hat{\mathbf{c}} \in \mathbb{R}^N$ in(5) can be rewritten as

$$\hat{\mathbf{c}} = (\mathbf{A}_\lambda^\top \mathbf{A}_\lambda)^{-1}\mathbf{A}_\lambda^\top \widetilde{\mathbf{y}} = \mathbf{A}_\lambda^\dagger \widetilde{\mathbf{y}}. \tag{9}$$

With the notations introduced above, we are ready to bound the Lipschitz constant of the trained randomized neural network.

Proof of Proposition 1. We first show that the coefficient is upper bounded. Specifically, we have

$$\|\hat{c}\|_2 = \|\mathbf{A}_\lambda^\dagger \widetilde{\mathbf{y}}\|_2 \le \|\mathbf{A}_\lambda^\dagger\|_2 \|\widetilde{\mathbf{y}}\|_2 \le \|(\sqrt{\lambda KM}\mathbf{W})^{-1}\|_2 \le \frac{1}{\sqrt{\lambda KM}\min_{k\in[N]}\|\boldsymbol{\omega}_k\|_2}$$

Then we can show that the Lipschitz constant is bounded by

$$\begin{aligned}\sum_{k=1}^N |\hat{c}_k|\|\boldsymbol{\omega}_k\| &\le \left(\sum_{k=1}^N |\hat{c}_k|^2\right)^{1/2}\left(\sum_{k=1}^N \|\boldsymbol{\omega}_k\|^2\right)^{1/2} \le \|\hat{c}\|_2\sqrt{N}\max_{k\in[N]}\|\boldsymbol{\omega}_k\|_2 \\ &\le \frac{\sqrt{N}\max_{k\in[N]}\|\boldsymbol{\omega}_k\|_2}{\sqrt{\lambda KM}\min_{k\in[N]}\|\boldsymbol{\omega}_k\|_2}.\end{aligned}$$

□

B Bernstein's Concentration Inequality

In this section, we recall two classical Bernstein's concentration inequalities from [19] that estimate the difference between empirical averages and true mean of random vectors. They are used to prove the generalization properties of randomized neural networks.

Theorem 5 (Vector-valued Bernstein's inequality in Hilbert space)**.** Let Z be an H-valued random variable, where $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$ is a separable Hilbert space. Suppose there exist positive numbers $b > 0$ and $c > 0$ such that

$$\mathbb{E}\|Z - \mathbb{E}Z\|^p \le \frac{1}{2}p!\sigma^2 b^{p-2} \text{ for all } p \ge 2. \tag{10}$$

For any $\delta \in (0,1)$ and $N \in \mathbb{N}$, denoting by $\{Z_n\}_{n=1}^N$ a sequence of N iid copies of Z, it holds that

$$\mathbb{P}\left\{\left\|\frac{1}{N}\sum_{n=1}^N Z_n - \mathbb{E}Z\right\| \le \frac{2b\log(2/\delta)}{N} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{N}}\right\} \ge 1 - \delta.$$

Lemma 6. Let Z be a uncentered random variable such that

$$\|Z\| \le c \text{ almost surely} \quad \text{and } \mathbb{E}\|Z - \mathbb{E}Z\|^2 \le \nu^2 \tag{11}$$

for some $c > 0$ and $\nu > 0$. Then Z satisfies Bernstein's moment condition (10) with $b = 2c$ and $\sigma = \nu$. If $\mathbb{E}Z = 0$, then taking $b = c$ suffices.

C Proof of Theorem 4

In this section, we provide a complete proof of Theorem 4 that gives an upper bound of the error $\sup_{x\in D}|f(x)-\hat{f}(x)|$.

Theorem 7. Let f be a function from $\mathcal{F}(\rho)$. Suppose that the random feature map ϕ satisfies $|\phi(\mathbf{x},\boldsymbol{\omega})|\le 1$ for all $\mathbf{x}\in X$ and $\boldsymbol{\omega}\in\mathbb{R}^d$. Then for any $\delta\in(0,1)$, there exists $c_1^*,\dots,c_N^*$ so that the function

$$f^*(\mathbf{x})=\sum_{k=1}^N c_k^*\phi(\mathbf{x},\boldsymbol{\omega}_k) \tag{12}$$

satisfies

$$\left|f(\mathbf{x})-f^\sharp(\mathbf{x})\right|\le\frac{12\|f\|_\rho\log(2/\delta)}{\sqrt{N}}$$

with probability at least $1-\delta$ over $\boldsymbol{\omega}_1,\dots,\boldsymbol{\omega}_N$ drawn i.i.d from $\rho(\boldsymbol{\omega})$.

Proof. We first introduce notations $\alpha_{\le T}(\boldsymbol{\omega})=\alpha(\boldsymbol{\omega})\mathbb{1}_{|\alpha(\boldsymbol{\omega})|\le T}$ and $\alpha_{>T}=\alpha(\boldsymbol{\omega})-\alpha_{\le T}(\boldsymbol{\omega})$ for any $T>0$. Then we define

$$c_k^\sharp=\alpha_{\le T}(\boldsymbol{\omega}_k)\quad\text{ for all } k\in[N], \tag{13}$$

where $\boldsymbol{\omega}_k$'s are i.i.d samples following a probability distribution with density $\rho(\boldsymbol{\omega})$, and hence define $f^\sharp(\mathbf{x})$ in (12) using $c_k^\sharp$'s. We can show that

$$\mathbb{E}f^\sharp(\mathbf{x})=\mathbb{E}_{\boldsymbol{\omega}}\left[\alpha_{\le T}(\boldsymbol{\omega})\phi(\mathbf{x},\boldsymbol{\omega})\right].$$

By utilizing the triangle inequality, we decompose the error into two terms

$$\left|f(\mathbf{x})-f^\sharp(\mathbf{x})\right|\le\underbrace{\left|f(\mathbf{x})-\mathbb{E}f^\sharp(\mathbf{x})\right|}_{I_1}+\underbrace{\left|\mathbb{E}f^\sharp(\mathbf{x})-f^\sharp(\mathbf{x})\right|}_{I_2}. \tag{14}$$

We first bound term I_1. Recalling the definitions of f and $\alpha_{\le T}(\boldsymbol{\omega})$, we bound term I_1 as

$$\begin{aligned}\left|f(\mathbf{x})-\mathbb{E}f^\sharp(\mathbf{x})\right|^2=&\left|\mathbb{E}_{\boldsymbol{\omega}}\left[\alpha_{>T}(\boldsymbol{\omega})\phi(\mathbf{x},\boldsymbol{\omega})\right]\right|^2\le\mathbb{E}_{\boldsymbol{\omega}}\left[\alpha(\boldsymbol{\omega})\right]^2\mathbb{E}_{\boldsymbol{\omega}}\left[\mathbb{1}_{|\alpha(\boldsymbol{\omega})|>T}\phi(\mathbf{x},\boldsymbol{\omega})\right]^2\\ =&\mathbb{E}_{\boldsymbol{\omega}}\left[\alpha(\boldsymbol{\omega})\right]^2\mathbb{P}\big(\alpha(\boldsymbol{\omega})^2>T^2\big)\le\frac{\Big(\mathbb{E}_{\boldsymbol{\omega}}[\alpha(\boldsymbol{\omega})^2]\Big)^2}{T^2}=\frac{\|f\|_\rho^4}{T^2}\end{aligned} \tag{15}$$

where we use the Cauchy-Schwarz inequality in the first line and the Markov's inequality in the second line.

Next, we bound term I_2. For any $\mathbf{x}\in X$, we define random variable $Z(\boldsymbol{\omega})=\alpha_{\le T}(\boldsymbol{\omega})\phi(\mathbf{x},\boldsymbol{\omega})$ and let $Z_1,\dots,Z_N$ be N i.i.d copies of Z defined by $Z_k=Z(\boldsymbol{\omega}_k)$ for each $k\in[N]$. By boundedness of $\alpha_{\le T}(\boldsymbol{\omega})$, we have an upper bound $|Z_k|\le T$ for any $k\in[N]$. The variance of Z is bounded above as

$$\sigma^2:=\mathbb{E}_{\boldsymbol{\omega}}|Z-\mathbb{E}_{\boldsymbol{\omega}}Z|^2\le\mathbb{E}_{\boldsymbol{\omega}}|Z|^2\le\mathbb{E}_{\boldsymbol{\omega}}[\alpha(\boldsymbol{\omega})^2]=\|f\|_\rho^2.$$

By Lemma 6 and Theorem 5, it holds that, with probability at least $1-\delta$,

$$\left|f^\sharp(\mathbf{x})-\mathbb{E}f^\sharp(\mathbf{x})\right|=\left|\frac{1}{N}\sum_{k=1}^N Z_k-\mathbb{E}_{\boldsymbol{\omega}}Z\right|\le\frac{4T\log(2/\delta)}{N}+\sqrt{\frac{2\|f\|_\rho^2\log(2/\delta)}{N}}. \tag{16}$$

Taking the square root for both sides of (15), and then adding it to (16) gives

$$|f(\mathbf{x})-f^\sharp(\mathbf{x})|\le\frac{\Big(\mathbb{E}_{\boldsymbol{\omega}}[\alpha(\boldsymbol{\omega})^2]\Big)}{T}+\frac{4T\log(2/\delta)}{N}+\sqrt{\frac{2\|f\|_\rho^2\log(2/\delta)}{N}}.$$

Selecting $T=\sqrt{N}\|f\|_\rho$ gives the desired result. □

Lemma 8 (Decay Rate of $\|\mathbf{c}^*\|_2^2$). The coefficient vector $\mathbf{c}^*$ of f^* holds, with probability at least $1-\delta$, that

$$\|\mathbf{c}^*\|_2\le\frac{12\|f\|_\rho\log(2/\delta)}{\sqrt{N}}$$

for any $\delta\in(0,1)$.

Proof. By the definition of c_k^*, we have

$$\|\mathbf{c}^*\|_2^2 = \sum_{k=1}^{N} |c_k^*|^2 = \frac{1}{N^2}\sum_{k=1}^{N} |\alpha_{\leq T}(\boldsymbol{\omega}_k)|^2. \tag{17}$$

Define random variable $Z(\boldsymbol{\omega}) = |\alpha_{\leq T}(\omega)|^2$ and let $Z_1, \dots, Z_N$ be N i.i.d copies of Z defined as $Z_k = |\alpha_{\leq T}(\boldsymbol{\omega}_k)|^2$ for each $k \in [N]$. By the boundedness of $\alpha_{\leq T}(\boldsymbol{\omega})$, we have an upper bound $|Z_k| \leq T^2$ for each $k \in [N]$. The variance of Z is bounded above as

$$\sigma^2 := \mathbb{E}_{\boldsymbol{\omega}}|Z - \mathbb{E}_{\boldsymbol{\omega}} Z|^2 \leq \mathbb{E}_{\boldsymbol{\omega}}|Z|^2 \leq T^2 \mathbb{E}_{\boldsymbol{\omega}}\left[|\alpha(\boldsymbol{\omega})|^2\right] = T^2\|f\|_\rho^2.$$

By Lemma 6 and Theorem 5, it holds with probability at least $1-\delta$ that

$$\left|\frac{1}{N}\sum_{k=1}^{N} Z_k - \mathbb{E}_{\boldsymbol{\omega}} Z\right| \leq \frac{4T^2\log(2/\delta)}{N} + \sqrt{\frac{2T^2\|f\|_\rho^2\log(2/\delta)}{N}}.$$

Then, we have

$$\|\mathbf{c}^*\|_2^2 = \frac{1}{N}\left(\frac{1}{N}\sum_{k=1}^{N}|\alpha_{\leq T}(\boldsymbol{\omega}_k)|^2\right) \leq \frac{1}{N}\left(\|f\|_\rho^2 + \frac{4T^2\log(2/\delta)}{N} + \sqrt{\frac{2T^2\|f\|_\rho^2\log(2/\delta)}{N}}\right).$$

Setting $T = \sqrt{N}\|f\|_\rho$ and taking the square root for both side give the desired result. □

Proof of Theorem 4. For each $x \in X$, we use triangluar inequality to obtain

$$|f(x) - \hat{f}(x)| \leq |f(x) - f^*(x)| + |f^*(x) - \hat{f}(x)|. \tag{18}$$

The bound of the first term $|f(x) - f^*(x)|$ is directly obtained by applying the result of Theorem 7, which is

$$|f(x) - f^*(x)| \leq \frac{12\|f\|_\rho \log(2/\delta)}{\sqrt{N}}. \tag{19}$$

Then we bound the second term $|f^*(x) - \hat{f}(x)|$.

$$|f^*(x) - \hat{f}(x)| = \left|\sum_{k=1}^{N} c_k^* - \hat{c}_k\phi(\langle\boldsymbol{\omega}_k, x\rangle + b_k)\right| \leq \sqrt{N}\|\mathbf{c}^* - \hat{\mathbf{c}}\|_2.$$

It remains to bound $\|\mathbf{c}^* - \hat{\mathbf{c}}\|_2$. Specifically, we have

$$\|\mathbf{c}^* - \hat{\mathbf{c}}\|_2^2 = \|\mathbf{A}_\lambda^\dagger\mathbf{A}_\lambda\mathbf{c}^* - \mathbf{A}_\lambda^\dagger\widetilde{\mathbf{y}}\|^2 \leq \|\mathbf{A}_\lambda^\dagger\|_2^2\left(\|\mathbf{A}\mathbf{c}^* - \mathbf{y}\|_2^2 + \lambda KM\|\mathbf{W}\mathbf{c}^*\|_2^2\right).$$

The (squared) operator norm of $\mathbf{A}_\lambda^\dagger$ is bounded by

$$\|\mathbf{A}_\lambda^\dagger\|_2^2 \leq \|(\sqrt{\lambda KM}\mathbf{W})^{-1}\|_2^2 \leq \frac{1}{\lambda KM\min_{k\in[N]}\|\boldsymbol{\omega}_k\|_2^2}. \tag{20}$$

Moreover, we can show that

$$\|\mathbf{A}\mathbf{c}^* - \mathbf{y}\|_2^2 = \sum_{j=1}^{KM}|f(\mathbf{x}_j) - f^*(\mathbf{x}_j)|^2 \leq KM\sup_{\mathbf{x}\in D}|f(\mathbf{x}) - f^*(\mathbf{x})|^2 \leq KM\left(\frac{12\|f\|_\rho\log(2/\delta)}{\sqrt{N}}\right)^2, \tag{21}$$

where we use Theorem 7 to obtain the last inequality. Then we apply Lemma 8 to show that

$$\|\mathbf{W}\mathbf{c}^*\|_2^2 \leq \|\mathbf{W}\|_2^2\|\mathbf{c}^*\|_2^2 \leq \max_{k\in[N]}\|\boldsymbol{\omega}_k\|_2^2\left(\frac{12\|f\|_\rho\log(2/\delta)}{\sqrt{N}}\right)^2. \tag{22}$$

Combining inequalities (20), (21), and (22) gives the following upper bound

$$|f^*(x) - \hat{f}(x)| \leq \left(\frac{1}{\sqrt{\lambda}\min_{k\in[N]}\|\boldsymbol{\omega}_k\|_2} + \frac{\max_{k\in[N]}\|\boldsymbol{\omega}_k\|_2}{\min_{k\in[N]}\|\boldsymbol{\omega}_k\|_2}\right)12\|f\|_\rho\log(2/\delta) \tag{23}$$

Adding inequalities 19 and 23 together leads to the desired result. □

D Asymptotic Analysis for the Lipschitz Constant

In this section, we numerically verify the asymptotic if the Lipschitz constant as we set $N, d \to \infty$. We consider the standard Gaussian and the uniform random weights, which are widely used in the training of randomized neural networks. We depict the results in Figure 4. Each point represents the average of 5 experiments. We observe that the ratio decreases as $d \to \infty$ and the ratio is not large when N is large.

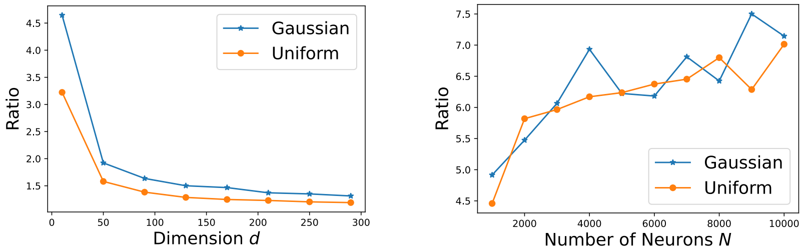
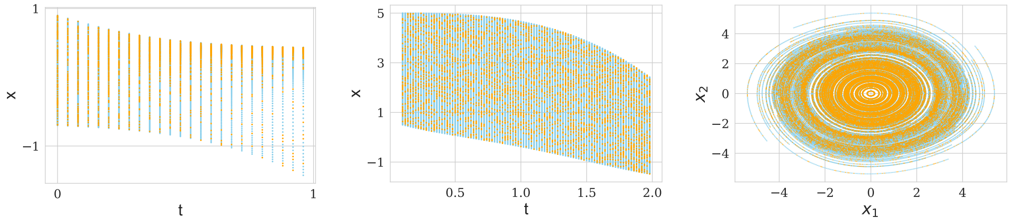


Figure 4: Asymptotic behaviors of the ratio. On the left: dimension d increases. On the right: the number of neurons N increases.

E More Experiment Results

E.1 Training and Test Samples

In Figure 5, we visualize training (in blue) and test (in orange) samples for all examples.



(a) Autonomous ODE

(b) Non-autonomous ODE

(c) Van der Pol Oscillator

Figure 5: Visualization of training (in blue) and test (in orange) data for the simulation studies.

E.2 Hyper-parameter Selection

In this section, we present more experiment results on the hyper-parameter selection. We fix the number of hidden neurons $N = 50$ and the variance of Gaussian random weights $\sigma^2 = 1$. The aim is to empirically verify the impact of regularization parameter λ. We report the numerical results of Non-autonomous ODE and Van der Pol Oscillator in Table 4 and Table 5, respectively. We observe that the regularized models have better generalization performance and control the Lipschitz constant as our theory shows. Moreover, the example of Van der Pol Oscillator has more samples resulting in smaller Lipschitz constant. This observation verifies the theory that the Lipschitz constant decreases as we increase the number of trajectories K and the number of time stamps M of each trajectory.

E.3 Solution Estimation

In this section, we present more empirical results on the solution estimation. Following the same experiment setup as Section 4.4, we use the trained randomized neural network as the right-hand side and numerically solve for the ODE system using the function odeint from the scipy package in Python. Instead of considering the noiseless regime, we consider the noisy setting here (1% and 5%).

We depict the true and predicted solutions of three ODE benchmarks in Figures 6 and 7. In the low noise regime (1%), the predicted solutions are accurate for long time prediction. In the high noise

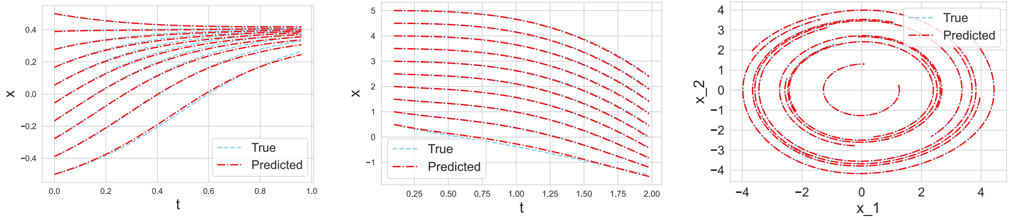
Regularization Parameter	Test Error	Generalization Gap	Recovery Error on Full Domain	Lipschitz Constant
0	1.95×10^{-1}	7.59×10^{-2}	3.75%	45469287.2
0.0001	$\mathbf{7.07 \times 10^{-3}}$	8.66×10^{-5}	**0.85%**	399.6
0.001	7.14×10^{-3}	7.01×10^{-5}	0.91%	222.3
0.01	7.38×10^{-3}	$\mathbf{4.43 \times 10^{-5}}$	0.95%	**116.8**

Table 4: Numerical results of Non-autonomous ODE: we report regularization parameters, test errors, generalization gaps, recovery errors on full domain, and Lipschitz constant of randomized neural networks in the noiseless regime.

Regularization Parameter	Test Error	Generalization Gap	Recovery Error on Full Domain	Lipschitz Constant
0	$\mathbf{2.69 \times 10^{-3}}$	-4.28×10^{-6}	0.23% 0.28%	95.18 196.32
0.0001	2.70×10^{-3}	-4.02×10^{-6}	**0.21%** **0.27%**	46.81 86.11
0.001	2.72×10^{-3}	$\mathbf{-7.24 \times 10^{-8}}$	0.25% 0.31%	36.87 60.14
0.01	2.80×10^{-3}	9.43×10^{-6}	0.33% 0.44%	**29.86** **43.04**

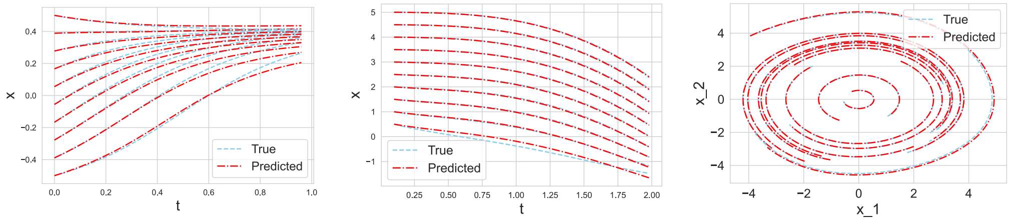
Table 5: Numerical results of Van der Pol Oscillator: we report regularization parameters, test errors, generalization gaps, recovery errors on full domain, and Lipschitz constant of randomized neural networks in the noiseless regime.

regime (5%), the predicted solutions are accurate for a short time prediction, but they are less accurate at the end of time interval. In summary, our results indicate that the predicted solutions are accurate. Therefore, our proposed regularized randomized neural networks are robust in noisy regimes.



(a) Autonomous ODE (b) Non-autonomous ODE (c) Van der Pol Oscillator

Figure 6: True and predicted solution of three ODE benchmarks. The noise level is 1%.



(a) Autonomous ODE (b) Non-autonomous ODE (c) Van der Pol Oscillator

Figure 7: True and predicted solution of three ODE benchmarks. The noise level is 5%.