Appendix

Organization of the Appendix Section A contains the proofs of the results of the main paper. Section B contains the details of the numerical illustrations presented in Section 4.3.

A Proofs

A.1 Proof of Proposition 1

The function

$$(t, h) \mapsto \sum_{i=1}^{m} \theta_i(t) f_i(h)$$

is locally Lipschitz-continuous with respect to its first variable and globally Lipschitz-continuous with respect to its second variable. Therefore the existence and uniqueness of the solution of the initial value problem (5) for $t \geqslant 0$ comes as a consequence of the Picard-Lindelöf theorem (see, e.g., Luk, 2017 for a self-contained presentation and Arnold, 1992 for a textbook).

A.2 Proof of Proposition 2

For $x \in \mathcal{X}$, let H be the solution of the initial value problem (5) with parameter θ and with the initial condition $H_0 = x$. Let us first upper-bound $\|f_i(H_t)\|$ for all $i \in \{1, \dots, m\}$ and $t > 0$. To this aim, for $t \geqslant 0$, we have

$$\begin{aligned}
\|H_t - H_0\| &= \Big\| \int_0^t \sum_{i=1}^{m} \theta_i(s) f_i(H_s) ds \Big\| \\
&\leqslant \int_0^t \sum_{i=1}^{m} |\theta_i(s)| \|f_i(H_0)\| ds + \int_0^t \sum_{i=1}^{m} |\theta_i(s)| \|f_i(H_s) - f_i(H_0)\| ds \\
&\leqslant M \int_0^t \sum_{i=1}^{m} |\theta_i(s)| ds + K_f \int_0^t \Big(\|H_s - H_0\| \sum_{i=1}^{m} |\theta_i(s)| \Big) ds \\
&\leqslant t M R_\Theta + K_f R_\Theta \int_0^t \|H_s - H_0\| ds.
\end{aligned}$$

Next, Grönwall's inequality yields, for $t \in [0, 1]$,

$$\|H_t - H_0\| \leqslant t M R_\Theta \exp(t K_f R_\Theta) \leqslant M R_\Theta \exp(K_f R_\Theta).$$

Hence

$$\|H_t\| \leqslant \|H_0\| + \|H_t - H_0\| \leqslant R_\mathcal{X} + M R_\Theta \exp(K_f R_\Theta),$$

yielding the first result of the proposition. Furthermore, for any $i \in \{1, \dots, m\}$,

$$\|f_i(H_t)\| \leqslant \|f_i(H_t) - f_i(H_0)\| + \|f_i(H_0)\| \leqslant M\big(K_f R_\Theta \exp(K_f R_\Theta) + 1\big) =: C.$$

Now, let $\tilde{H}$ be the solution of the initial value problem (5) with another parameter $\tilde{\theta}$ and with the same initial condition $\tilde{H}_0 = x$. Then, for any $t \geqslant 0$,

$$H_t - \tilde{H}_t = \int_0^t \sum_{i=1}^{m} \theta_i(s) f_i(H_s) ds - \int_0^t \sum_{i=1}^{m} \tilde{\theta}_i(s) f_i(\tilde{H}_s) ds.$$

Hence

$$\begin{aligned}\|H_t - \tilde{H}_t\| &= \Big\| \int_0^t \sum_{i=1}^m (\theta_i(s) - \tilde{\theta}_i(s)) f_i(H_s) ds + \int_0^t \sum_{i=1}^m \tilde{\theta}_i(s)(f_i(H_s) - f_i(\tilde{H}_s)) ds \Big\| \\ &\leqslant \int_0^t \sum_{i=1}^m |\theta_i(s) - \tilde{\theta}_i(s)| \|f_i(H_s)\| ds + \int_0^t \sum_{i=1}^m |\tilde{\theta}_i(s)| \|f_i(H_s) - f_i(\tilde{H}_s)\| ds \\ &\leqslant \int_0^t \sum_{i=1}^m |\theta_i(s) - \tilde{\theta}_i(s)| \|f_i(H_s)\| ds + K_f \int_0^t \Big(\|H_s - \tilde{H}_s\| \sum_{i=1}^m |\tilde{\theta}_i(s)| \Big) ds \\ &\leqslant tC\|\theta - \tilde{\theta}\|_{1,\infty} + K_f R_\Theta \int_0^t \|H_s - \tilde{H}_s\| ds.\end{aligned}$$

Then Grönwall's inequality implies that, for $t \in [0,1]$,

$$\begin{aligned}\|H_t - \tilde{H}_t\| &\leqslant tC\|\theta - \tilde{\theta}\|_{1,\infty} \exp(tK_f R_\Theta) \\ &\leqslant M(K_f R_\Theta \exp(K_f R_\Theta) + 1) \exp(K_f R_\Theta) \|\theta - \tilde{\theta}\|_{1,\infty} \\ &\leqslant 2MK_f R_\Theta \exp(2K_f R_\Theta) \|\theta - \tilde{\theta}\|_{1,\infty}\end{aligned}$$

since $1 \leqslant K_f R_\Theta \exp(K_f R_\Theta)$ because $K_f \geqslant 1$, $R_\Theta \geqslant 1$.

A.3 Proof of Proposition 3

We first prove the result for $m = 1$. Let G_x be an $\varepsilon/2K_\Theta$-grid of $[0,1]$ and G_y an $\varepsilon/2$-grid of $[-R_\Theta, R_\Theta]$. Formally, we can take

$$G_x = \Big\{ \frac{k\varepsilon}{2K_\Theta},\ 0 \leqslant k \leqslant \Big\lceil \frac{2K_\Theta}{\varepsilon} \Big\rceil \Big\} \quad \text{and} \quad G_y = \Big\{ -R_\Theta + \frac{k\varepsilon}{2},\ 1 \leqslant k \leqslant \Big\lfloor \frac{4R_\Theta}{\varepsilon} \Big\rfloor \Big\}$$

Our cover consists of all functions that start at a point of G_y, are piecewise linear with kinks in G_x, where each piece has slope $+K_\Theta$ or $-K_\Theta$. Hence our cover is of size

$$\mathcal{N}_1(\varepsilon) = |G_y| 2^{|G_x|} \leqslant \frac{4R_\Theta}{\varepsilon} 2^{\frac{2K_\Theta}{\varepsilon}+2} = \frac{16R_\Theta}{\varepsilon} 4^{\frac{K_\Theta}{\varepsilon}}.$$

Now take a function $f : [0,1] \to \mathbb{R}$ that is uniformly bounded by R_Θ and K_Θ-Lipschitz. We construct a cover member at distance ε from f as follows. Choose a point y_0 in G_y at distance at most $\varepsilon/2$ from $f(0)$. Since $f(0) \in [-R_\Theta, R_\Theta]$, this is clearly possible, except perhaps at the end of the interval. To verify that it is possible at the end of the interval, note that R_Θ is at a distance less than $\varepsilon/2$ of the last element of the grid, since

$$R_\Theta - \Big(-R_\Theta + \Big\lfloor \frac{4R_\Theta}{\varepsilon} \Big\rfloor \frac{\varepsilon}{2} \Big) = 2R_\Theta - \Big\lfloor \frac{4R_\Theta}{\varepsilon} \Big\rfloor \frac{\varepsilon}{2} \in \Big[2R_\Theta - \frac{4R_\Theta}{\varepsilon}\frac{\varepsilon}{2}, 2R_\Theta - \Big(\frac{4R_\Theta}{\varepsilon} - 1 \Big) \frac{\varepsilon}{2} \Big] = \Big[0, \frac{\varepsilon}{2} \Big].$$

Then, among the cover members that start at y_0, choose the one which is closest to f at each point of G_x (in case of equality, pick any one). Let us denote this cover member as $\tilde{f}$. Let us show recursively that f is at ℓ_∞-distance at most ε from $\tilde{f}$. More precisely, let us first show by induction on k that for all $k \in \{0, \dots, \lceil \frac{2K_\Theta}{\varepsilon} \rceil\}$,

$$\big| f\big(\frac{k\varepsilon}{2K_\Theta}\big) - \tilde{f}\big(\frac{k\varepsilon}{2K_\Theta}\big) \big| \leqslant \frac{\varepsilon}{2}. \tag{15}$$

First, $|f(0) - \tilde{f}(0)| \leqslant \frac{\varepsilon}{2}$. Then, assume that (15) holds for some k. Then we have the following inequalities:

$$\begin{aligned}\tilde{f}\big(\frac{k\varepsilon}{2K_\Theta}\big) - \varepsilon &\leqslant f\big(\frac{k\varepsilon}{2K_\Theta}\big) - \frac{\varepsilon}{2} && \text{(by induction)} \\ &\leqslant f\big(\frac{(k+1)\varepsilon}{2K_\Theta}\big) && (f \text{ is } K_\Theta\text{-Lipschitz}) \\ &\leqslant f\big(\frac{k\varepsilon}{2K_\Theta}\big) + \frac{\varepsilon}{2} && (f \text{ is } K_\Theta\text{-Lipschitz}) \\ &\leqslant \tilde{f}\big(\frac{k\varepsilon}{2K_\Theta}\big) + \varepsilon && \text{(by induction).}\end{aligned}$$

Moreover, by definition, $\tilde{f}\big(\frac{(k+1)\varepsilon}{K_\Theta}\big)$ is the closest point to $f\big(\frac{(k+1)\varepsilon}{K_\Theta}\big)$ among

$$\Big\{\tilde{f}\big(\frac{k\varepsilon}{K_\Theta}\big) - \frac{\varepsilon}{2}, \tilde{f}\big(\frac{k\varepsilon}{K_\Theta}\big) + \frac{\varepsilon}{2}\Big\}.$$

The bounds above show that, among those two points, at least one is at distance no more than $\varepsilon/2$ from $f\big(\frac{(k+1)\varepsilon}{K_\Theta}\big)$. This shows (15) at rank $k+1$.

To conclude, take now $x \in [0,1]$. There exists $k \in \{0, \dots, \lceil \frac{2K_\Theta}{\varepsilon} \rceil\}$ such that x is at distance at most $\varepsilon/4K_\Theta$ from $\frac{k\varepsilon}{2K_\Theta}$. Again, this is clear except perhaps at the end of the interval, where it is also true since

$$1 - \Big\lceil \frac{2K_\Theta}{\varepsilon} \Big\rceil \frac{\varepsilon}{2K_\Theta} \leqslant 1 - \frac{2K_\Theta}{\varepsilon}\frac{\varepsilon}{2K_\Theta} = 0,$$

meaning that 1 is located between two elements of the grid G_x, showing that it is at distance at most $\varepsilon/4K_\Theta$ from one element of the grid. Then, we have

$$\begin{aligned}|f(x) - \tilde{f}(x)| &\leqslant \Big|f(x) - f\big(\frac{k\varepsilon}{2K_\Theta}\big)\Big| + \Big|f\big(\frac{k\varepsilon}{2K_\Theta}\big) - \tilde{f}\big(\frac{k\varepsilon}{2K_\Theta}\big)\Big| + \Big|\tilde{f}\big(\frac{k\varepsilon}{2K_\Theta}\big) - \tilde{f}(x)\Big| \\ &\leqslant \frac{\varepsilon}{4} + \frac{\varepsilon}{2} + \frac{\varepsilon}{4},\end{aligned}$$

where the first and third terms are upper-bounded because f and $\tilde{f}$ are K_Θ-Lip, while the second term is upper bounded by (15). Hence $\|f - \tilde{f}\|_\infty \leqslant \varepsilon$, proving the result for $m = 1$.

Finally, to prove the result for a general m, note that the Cartesian product of ε/m-covers for each coordinate of θ gives a ε-cover for θ. Indeed, consider such covers and take $\theta \in \Theta$. Since each coordinate of θ is uniformly bounded by R_Θ and K_Θ-Lipschitz, the proof above shows the existence of a cover member $\tilde{\theta}$ such that, for all $i \in \{1, \dots, m\}$, $\|\theta_i - \tilde{\theta}_i\|_\infty \leqslant \varepsilon/m$. Then

$$\|\theta - \tilde{\theta}\|_{1,\infty} = \sup_{0 \leqslant t \leqslant 1} \sum_{i=1}^m |\theta_i(t) - \tilde{\theta}_i(t)| \leqslant \sup_{0 \leqslant t \leqslant 1} \sum_{i=1}^m \|\theta_i - \tilde{\theta}_i\|_\infty \leqslant \varepsilon.$$

As a consequence, we conclude that

$$\mathcal{N}(\varepsilon) \leqslant \Big(\mathcal{N}_1\big(\frac{\varepsilon}{m}\big)\Big)^m = \Big(\frac{16mR_\Theta}{\varepsilon}\Big)^m 4^{\frac{m^2 K_\Theta}{\varepsilon}}.$$

Taking the logarithm yields the result.

A.4 Proof of Theorem 1

First note that, for any $\theta \in \Theta$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$|\ell(F_\theta(x), y)| \leqslant |\ell(F_\theta(x), y) - \ell(y, y)| + |\ell(y, y)| \leqslant K_\ell \|F_\theta(x) - y\|.$$

since, by assumption, ℓ is K_ℓ-Lipschitz with respect to its first variable and $\ell(y, y) = 0$. Thus

$$|\ell(F_\theta(x), y)| \leqslant K_\ell(\|F_\theta(x)\| + \|y\|) \leqslant K_\ell\big(R_\mathcal{X} + MR_\Theta \exp(K_f R_\Theta) + R_\mathcal{Y}\big) =: \overline{M}.$$

by Proposition 2.

Now, taking $\delta > 0$, a classical computation involving McDiarmid's inequality (see, e.g., Wainwright, 2019, proof of thm 4.10) yields that, with probability at least $1 - \delta$,

$$\mathscr{R}(\widehat{\theta}_n) \leqslant \widehat{\mathscr{R}}_n(\widehat{\theta}_n) + \mathbb{E}\Big[\sup_{\theta \in \Theta} |\mathscr{R}(\theta) - \widehat{\mathscr{R}}_n(\theta)|\Big] + \frac{\overline{M}\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}.$$

Denote $C = 2MK_f R_\Theta \exp(2K_f R_\Theta)$. Then we show that $\mathscr{R}$ and $\widehat{\mathscr{R}}_n$ are CK_ℓ-Lipschitz with respect to $(\theta, \|\cdot\|_{1,\infty})$: for $\theta, \tilde{\theta} \in \Theta$,

$$\begin{aligned}|\mathscr{R}(\theta) - \mathscr{R}(\tilde{\theta})| &\leqslant \mathbb{E}\big[|\ell(F_\theta(x), y) - \ell(F_{\tilde{\theta}}(x), y)|\big] \\ &\leqslant K_\ell \mathbb{E}\big[\|F_\theta(x) - F_{\tilde{\theta}}(x)\|\big] \\ &\leqslant CK_\ell \|\theta - \tilde{\theta}\|_{1,\infty},\end{aligned}$$

according to Proposition 2. The proof for the empirical risk is very similar.

Let now $\varepsilon > 0$ and $\mathcal{N}(\varepsilon)$ be the covering number of Θ endowed with the $(1, \infty)$-norm. By Proposition 3,

$$\log \mathcal{N}(\varepsilon) \leqslant m \log\Big(\frac{16 m R_\Theta}{\varepsilon}\Big) + \frac{m^2 K_\Theta \log(4)}{\varepsilon}.$$

Take $\theta^{(1)}, \dots, \theta^{(\mathcal{N}(\varepsilon))}$ the associated cover elements. Then, for any $\theta \in \Theta$, denoting $\theta^{(i)}$ the cover element at distance at most ε from θ,

$$\begin{aligned}|\mathscr{R}(\theta) - \widehat{\mathscr{R}}_n(\theta)| &\leqslant |\mathscr{R}(\theta) - \mathscr{R}(\theta^{(i)})| + |\mathscr{R}(\theta^{(i)}) - \widehat{\mathscr{R}}_n(\theta^{(i)})| + |\widehat{\mathscr{R}}_n(\theta^{(i)}) - \widehat{\mathscr{R}}_n(\theta)| \\ &\leqslant 2CK_\ell \varepsilon + \sup_{i \in \{1, \dots, \mathcal{N}(\varepsilon)\}} |\mathscr{R}(\theta^{(i)}) - \widehat{\mathscr{R}}_n(\theta^{(i)})|.\end{aligned}$$

Hence

$$\mathbb{E}\Big[\sup_{\theta \in \Theta} |\mathscr{R}(\theta) - \widehat{\mathscr{R}}_n(\theta)|\Big] \leqslant 2CK_\ell \varepsilon + \mathbb{E}\Big[\sup_{i \in \{1, \dots, \mathcal{N}(\varepsilon)\}} |\mathscr{R}(\theta^{(i)}) - \widehat{\mathscr{R}}_n(\theta^{(i)})|\Big].$$

Since $\widehat{\mathscr{R}}_n(\theta)$ is the average of n independent random variables, which are each almost surely bounded by $\overline{M}$, it is $\overline{M}/\sqrt{n}$ sub-Gaussian, hence we have the classical inequality on the expectation of the maximum of sub-Gaussian random variables (see, e.g., Rigollet and Hütter, 2017, Theorem 1.14)

$$\mathbb{E}\Big[\sup_{i \in \{1, \dots, \mathcal{N}(\varepsilon)\}} |\mathscr{R}(\theta^{(i)}) - \widehat{\mathscr{R}}_n(\theta^{(i)})|\Big] \leqslant \overline{M}\sqrt{\frac{2\log(2\mathcal{N}(\varepsilon))}{n}}.$$

The remainder of the proof consists in computations to put the result in the required format. More precisely, we have

$$\begin{aligned}\mathbb{E}\Big[\sup_{\theta \in \Theta} |\mathscr{R}(\theta) - \widehat{\mathscr{R}}_n(\theta)|\Big] &\leqslant 2CK_\ell \varepsilon + \overline{M}\sqrt{\frac{2\log(2\mathcal{N}(\varepsilon))}{n}} \\ &\leqslant 2CK_\ell \varepsilon + \overline{M}\sqrt{\frac{2\log(2) + 2m\log\big(\frac{16mR_\Theta}{\varepsilon}\big) + \frac{2m^2 K_\Theta}{\varepsilon}\log(4)}{n}} \\ &\leqslant 2CK_\ell \varepsilon + \overline{M}\sqrt{\frac{2(m+1)\log\big(\frac{16mR_\Theta}{\varepsilon}\big) + \frac{2m^2 K_\Theta}{\varepsilon}\log(4)}{n}}.\end{aligned}$$

The third step is valid if $\frac{16mR_\Theta}{\varepsilon} \geqslant 2$. We will shortly take ε to be equal to $\frac{1}{\sqrt{n}}$, thus this condition holds true under the assumption from the Theorem that $mR_\Theta\sqrt{n} \geqslant 3$. Hence we obtain

$$\mathscr{R}(\widehat{\theta}_n) \leqslant \widehat{\mathscr{R}}_n(\widehat{\theta}_n) + 2CK_\ell \varepsilon + \overline{M}\sqrt{\frac{2(m+1)\log\big(\frac{16mR_\Theta}{\varepsilon}\big) + \frac{2m^2 K_\Theta}{\varepsilon}\log(4)}{n}} + \frac{\overline{M}\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}. \tag{16}$$

Now denote $\tilde{B} = 2\overline{M}K_f \exp(K_f R_\Theta)$. Then $CK_\ell \leqslant \tilde{B}$ and $2\overline{M} \leqslant \tilde{B}$. Taking $\varepsilon = \frac{1}{\sqrt{n}}$, we obtain

$$\begin{aligned}\mathscr{R}(\widehat{\theta}_n) &\leqslant \widehat{\mathscr{R}}_n(\widehat{\theta}_n) + \frac{2\tilde{B}}{\sqrt{n}} + \frac{\tilde{B}}{2}\sqrt{\frac{2(m+1)\log(16mR_\Theta\sqrt{n})}{n} + \frac{2m^2 K_\Theta \log(4)}{\sqrt{n}}} + \frac{\tilde{B}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}} \\ &\leqslant \widehat{\mathscr{R}}_n(\widehat{\theta}_n) + \frac{2\tilde{B}}{\sqrt{n}} + \frac{\tilde{B}}{2}\sqrt{\frac{2(m+1)\log(16mR_\Theta\sqrt{n})}{n}} + \frac{\tilde{B}}{2}\frac{m\sqrt{2K_\Theta\log(4)}}{n^{1/4}} + \frac{\tilde{B}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}} \\ &\leqslant \widehat{\mathscr{R}}_n(\widehat{\theta}_n) + \frac{3\tilde{B}}{2}\sqrt{\frac{2(m+1)\log(16mR_\Theta\sqrt{n})}{n}} + \tilde{B}\frac{m\sqrt{K_\Theta}}{n^{1/4}} + \frac{\tilde{B}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}},\end{aligned}$$

since $2 \leqslant 2\sqrt{\log(2)} \leqslant \sqrt{2(m+1)\log(16mR_\Theta\sqrt{n})}$ since $16mR_\Theta\sqrt{n} \geqslant 2$ by the Theorem's assumptions, and $\sqrt{2\log(4)} \leqslant 2$. We finally obtain that

$$\mathscr{R}(\widehat{\theta}_n) \leqslant \widehat{\mathscr{R}}_n(\widehat{\theta}_n) + 3\tilde{B}\sqrt{\frac{(m+1)\log(mR_\Theta n)}{n}} + \tilde{B}\frac{m\sqrt{K_\Theta}}{n^{1/4}} + \frac{\tilde{B}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}},$$

by noting that $n \geqslant 9\max(m^{-2}R_\Theta^{-2}, 1)$ implies that

$$\log(16mR_\Theta\sqrt{n}) \leqslant 2\log(mR_\Theta n).$$

The result unfolds since the constant B in the Theorem is equal to $3\tilde{B}$.

A.5 Proof of Corollary 1

The corollary is an immediate consequence of Theorem 1. To obtain the result, note that $m = d^2$, thus in particular $\sqrt{m+1} = \sqrt{d^2+1} \leqslant d+1$, and besides $\log(R_{\mathcal{W}} d^2 n) \leqslant 2\log(R_{\mathcal{W}} dn)$ since $R_{\mathcal{W}} n \leqslant R_{\mathcal{W}}^2 n^2$ by assumption on n.

A.6 Proof of Proposition 4

For $x \in \mathcal{X}$, let $(H_k)_{0 \leqslant k \leqslant L}$ be the values of the layers defined by the recurrence (10) with the weights $\mathbf{W}$ and the input $H_0 = x$. We denote by $\|\cdot\|$ the ℓ_2-norm for vectors and the spectral norm for matrices. Then, for $k \in \{0, \dots, L-1\}$, we have

$$\|H_{k+1}\| \leqslant \|H_k\| + \frac{1}{L}\|W_k \sigma(H_k)\| \leqslant \|H_k\| + \frac{1}{L}\|W_k\| \, \|\sigma(H_k)\| \leqslant \Big(1 + \frac{K_\sigma R_{\mathcal{W}}}{L}\Big)\|H_k\|,$$

where the last inequality uses that the spectral norm of a matrix is upper-bounded by its $(1,1)$-norm and that $\sigma(0) = 0$. As a consequence, for any $k \in \{0, \dots, L\}$,

$$\|H_k\| \leqslant \Big(1 + \frac{K_\sigma R_{\mathcal{W}}}{L}\Big)^k \|H_0\| \leqslant \exp(K_\sigma R_{\mathcal{W}}) R_{\mathcal{X}} =: C,$$

yielding the first claim of the Proposition.

Now, let $\tilde{H}$ be the values of the layers (10) with another parameter $\tilde{\mathbf{W}}$ and with the same input $\tilde{H}_0 = x$. Then, for any $k \in \{0, \dots, L-1\}$,

$$H_{k+1} - \tilde{H}_{k+1} = H_k - \tilde{H}_k + \frac{1}{L}(W_k \sigma(H_k) - \tilde{W}_k \sigma(\tilde{H}_k)).$$

Hence, using again that the spectral norm of a matrix is upper-bounded by its $(1,1)$-norm and that $\sigma(0) = 0$,

$$\begin{aligned}
\|H_{k+1} - \tilde{H}_{k+1}\| &\leqslant \|H_k - \tilde{H}_k\| + \frac{1}{L}\|W_k(\sigma(H_k) - \sigma(\tilde{H}_k))\| + \frac{1}{L}\|(W_k - \tilde{W}_k)\sigma(\tilde{H}_k)\| \\
&\leqslant \Big(1 + K_\sigma \frac{R_{\mathcal{W}}}{L}\Big)\|H_k - \tilde{H}_k\| + \frac{K_\sigma}{L}\|W_k - \tilde{W}_k\| \, \|\tilde{H}_k\| \\
&\leqslant \Big(1 + K_\sigma \frac{R_{\mathcal{W}}}{L}\Big)\|H_k - \tilde{H}_k\| + \frac{C K_\sigma}{L}\|W_k - \tilde{W}_k\|.
\end{aligned}$$

Then, dividing by $(1 + K_\sigma \frac{R_{\mathcal{W}}}{L})^{k+1}$ and using the method of differences, we obtain that

$$\begin{aligned}
\frac{\|H_k - \tilde{H}_k\|}{(1 + K_\sigma \frac{R_{\mathcal{W}}}{L})^k} &\leqslant \|H_0 - \tilde{H}_0\| + \frac{C K_\sigma}{L} \sum_{j=0}^{k-1} \frac{\|W_j - \tilde{W}_j\|}{(1 + K_\sigma \frac{R_{\mathcal{W}}}{L})^{j+1}} \\
&\leqslant \frac{C K_\sigma}{L}\|\mathbf{W} - \tilde{\mathbf{W}}\|_{1,1,\infty} \sum_{j=0}^{k-1} \frac{1}{(1 + K_\sigma \frac{R_{\mathcal{W}}}{L})^{j+1}}.
\end{aligned}$$

Finally note that

$$\begin{aligned}
\sum_{j=0}^{k-1} \frac{(1 + K_\sigma \frac{R_{\mathcal{W}}}{L})^k}{(1 + K_\sigma \frac{R_{\mathcal{W}}}{L})^{j+1}} &= \sum_{j=0}^{k-1} (1 + K_\sigma \frac{R_{\mathcal{W}}}{L})^j \\
&= \frac{L}{K_\sigma R_{\mathcal{W}}}\Big((1 + K_\sigma \frac{R_{\mathcal{W}}}{L})^k - 1\Big) \\
&\leqslant \frac{L}{K_\sigma R_{\mathcal{W}}}(\exp(K_\sigma R_{\mathcal{W}}) - 1).
\end{aligned}$$

We conclude that

$$\|H_k - \tilde{H}_k\| \leqslant \frac{C}{R_{\mathcal{W}}}(\exp(K_\sigma R_{\mathcal{W}}) - 1)\|\mathbf{W} - \tilde{\mathbf{W}}\|_{1,1,\infty} \leqslant \frac{R_{\mathcal{X}}}{R_{\mathcal{W}}} \exp(2K_\sigma R_{\mathcal{W}})\|\mathbf{W} - \tilde{\mathbf{W}}\|_{1,1,\infty}.$$

A.7 Proof of Proposition 5

For two integers a and b, denote respectively $a//b$ and $a\%b$ the quotient and the remainder of the Euclidian division of a by b. Then, for $\mathbf{W} \in \mathbb{R}^{L\times d\times d}$, let $\phi(\mathbf{W}) : [0,1] \to \mathbb{R}^{d^2}$ the piecewise-affine function defined as follows: $\phi(\mathbf{W})$ is affine on every interval $\left[\frac{k}{L}, \frac{k+1}{L}\right]$ for $k \in \{0,\dots,L-1\}$; for $k \in \{1,\dots,L\}$ and $i \in \{1,\dots,d^2\}$,

$$\phi(\mathbf{W})_i\Big(\frac{k}{L}\Big) = \mathbf{W}_{\frac{k}{L},(i//d)+1,(i\%d)+1}\,,$$

and $\phi(\mathbf{W})_i(0) = \phi(\mathbf{W})_i(1/L)$. Then $\phi(\mathbf{W})$ satisfies two properties. First, it is a linear function of $\mathbf{W}$. Second, for $\mathbf{W} \in \mathbb{R}^{L\times d\times d}$,

$$\|\phi(\mathbf{W})\|_{1,\infty} = \|\mathbf{W}\|_{1,1,\infty},$$

because, for $x \in [0,1]$, $\phi(\mathbf{W})(x)$ is a convex combination of two vectors that are bounded in ℓ_1-norm by $\|\mathbf{W}\|_{1,1,\infty}$, so it is itself bounded in ℓ_1-norm by $\|\mathbf{W}\|_{1,1,\infty}$, implying that $\|\phi(\mathbf{W})\|_{1,\infty} \leqslant \|\mathbf{W}\|_{1,1,\infty}$. Reciprocally,

$$\|\phi(\mathbf{W})\|_{1,\infty} = \sup_{0\leqslant t\leqslant 1} \|\phi(\mathbf{W})(x)\|_1 \geqslant \sup_{1\leqslant k\leqslant L} \Big\|\phi(\mathbf{W})\Big(\frac{k}{L}\Big)\Big\|_1 = \|\mathbf{W}\|_{1,1,\infty}.$$

Now, take $\mathbf{W} \in \mathcal{W}$. The second property of ϕ implies that $\|\phi(\mathbf{W})\|_{1,\infty} \leqslant R_{\mathcal{W}}$. Moreover, each coordinate of $\phi(\mathbf{W})$ is $K_{\mathcal{W}}$-Lipschitz, since the slope of each piece of $\phi(\mathbf{W})_i$ is at most $K_{\mathcal{W}}$. As a consequence, $\phi(\mathbf{W})$ belongs to

$$\Theta_{\mathcal{W}} = \{\theta : [0,1] \to \mathbb{R}^{d^2},\ \|\theta\|_{1,\infty} \leqslant R_{\mathcal{W}} \text{ and } \theta_i \text{ is } K_{\mathcal{W}}\text{-Lipschitz for } i \in \{1,\dots,d^2\}\}.$$

Therefore $\phi(\mathcal{W})$ is a subset of $\Theta_{\mathcal{W}}$, thus its covering number is less than the one of $\Theta_{\mathcal{W}}$. Moreover, ϕ is clearly injective, thus we can define ϕ^{-1} on its image. Consider an ε-cover $(\theta_1,\dots,\theta_N)$ of $(\phi(\mathcal{W}), \|\cdot\|_{1,\infty})$. Let us show that $(\phi^{-1}(\theta_1),\dots,\phi^{-1}(\theta_N))$ is an an ε-cover of $(\mathcal{W}, \|\cdot\|_{1,1,\infty})$: take $\mathbf{W} \in \mathcal{W}$ and consider θ_i a cover member at distance less than ε from $\phi(\mathbf{W})$. Then

$$\|\mathbf{W} - \phi^{-1}(\theta_i)\|_{1,1,\infty} = \|\phi(\mathbf{W} - \phi^{-1}(\theta_i))\|_{1,\infty} = \|\phi(\mathbf{W}) - \theta_i\|_{1,\infty} \leqslant \varepsilon,$$

where the second equality holds by linearity of ϕ. Therefore, the covering number of $(\mathcal{W}, \|\cdot\|_{1,1,\infty})$ is upper bounded by the one of $(\phi(\mathcal{W}), \|\cdot\|_{1,\infty})$, which itself is upper bounded by the one of $(\Theta_{\mathcal{W}}, \|\cdot\|_{1,\infty})$, yielding the result by Proposition 3.

A.8 Proof of Theorem 2

The proof structure is the same as the one of Theorem 1, but some constants change. Similarly to (16), we obtain that, if $\frac{16d^2R_{\mathcal{W}}}{\varepsilon} \geqslant 2$ (which holds true for $\varepsilon = 1/\sqrt{n}$ and under the assumption of the Theorem),

$$\mathscr{R}(\widehat{\mathbf{W}}_n) \leqslant \widehat{\mathscr{R}}_n(\widehat{\mathbf{W}}_n) + 2CK_\ell\varepsilon + \overline{M}\sqrt{\frac{2(d^2+1)\log\left(\frac{16d^2R_{\mathcal{W}}}{\varepsilon}\right) + \frac{2d^4K_{\mathcal{W}}}{\varepsilon}\log(4)}{n}} + \frac{\overline{M}\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}},$$

with

$$\overline{M} = K_\ell(R_{\mathcal{X}}\exp(K_\sigma R_{\mathcal{W}}) + R_{\mathcal{Y}})$$

and

$$C = \frac{R_{\mathcal{X}}}{R_{\mathcal{W}}}\exp(2K_\sigma R_{\mathcal{W}}).$$

Finally denote

$$\tilde{B} = 2\overline{M}\max\Big(\frac{\exp(K_\sigma R_{\mathcal{W}})}{R_{\mathcal{W}}}, 1\Big).$$

Then $CK_\ell \leqslant \tilde{B}$ and $2\overline{M} \leqslant \tilde{B}$. Taking $\varepsilon = \frac{1}{\sqrt{n}}$, we obtain as in the proof of Theorem 1 that

$$\mathscr{R}(\widehat{\mathbf{W}}_n) \leqslant \widehat{\mathscr{R}}_n(\widehat{\mathbf{W}}_n) + 3\tilde{B}\sqrt{\frac{(d^2+1)\log(d^2R_{\mathcal{W}}n)}{n}} + \tilde{B}\frac{d^2\sqrt{K_{\mathcal{W}}}}{n^{1/4}} + \frac{\tilde{B}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}.$$

for $n \geqslant 9R_{\mathcal{W}}^{-1}\max(d^{-4}R_{\mathcal{W}}^{-1}, 1)$. Thus

$$\mathscr{R}(\widehat{\mathbf{W}}_n) \leqslant \widehat{\mathscr{R}}_n(\widehat{\mathbf{W}}_n) + 3\sqrt{2}\tilde{B}(d+1)\sqrt{\frac{\log(dR_{\mathcal{W}}n)}{n}} + \tilde{B}\frac{d^2\sqrt{K_{\mathcal{W}}}}{n^{1/4}} + \frac{\tilde{B}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}},$$

since $\sqrt{d^2+1} \leqslant d+1$ and $R_{\mathcal{W}}n \leqslant R_{\mathcal{W}}^2 n^2$ by assumption on n. The result unfolds since the constant B in the Theorem is equal to $3\sqrt{2}\tilde{B}$.

A.9 Proof of Corollary 2

Let

$$A(\mathbf{W}) = \Big(\prod_{k=1}^{L}\Big\|I + \frac{1}{L}W_k\Big\|\Big)\Big(\sum_{k=1}^{L}\frac{\|W_k^T\|_{2,1}^{2/3}}{L^{2/3}\|I+\frac{1}{L}W_k\|^{2/3}}\Big)^{3/2},$$

where $\|\cdot\|_{2,1}$ denotes the $(2,1)$-norm defined as the ℓ_1-norm of the ℓ_2-norms of the columns, and I is the identity matrix (and we recall that $\|\cdot\|$ denotes the spectral norm). We apply Theorem 1.1 from Bartlett et al. (2017) by taking as reference matrices the identity matrix. The theorem shows that, under the assumptions of the corollary,

$$\mathbb{P}\Big(\arg\max_{1\leqslant j\leqslant d} F_{\mathbf{W}}(x)_j \neq y\Big) \leqslant \widehat{\mathscr{R}}_n(\mathbf{W}) + C\frac{R_{\mathcal{X}}A(\mathbf{W})\log(d)}{\gamma\sqrt{n}} + \frac{C}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}},$$

where, as in the corollary, $\widehat{\mathscr{R}}_n(\mathbf{W}) \leqslant n^{-1}\sum_{i=1}^{n}\mathbf{1}_{F_{\mathbf{W}}(x_i)_{y_i} \leqslant \gamma + \max_{j\neq y_i} f(x_i)_j}$ and C is a universal constant. Let us upper bound $A(\mathbf{W})$ to conclude. On the one hand, we have

$$\begin{aligned}
\prod_{k=1}^{L}\Big\|I + \frac{1}{L}W_k\Big\| &\leqslant \prod_{k=1}^{L}\Big(\|I\| + \frac{1}{L}\|W_k\|\Big)\\
&\leqslant \prod_{k=1}^{L}\Big(1 + \frac{1}{L}\|W_k\|_{1,1}\Big)\\
&\leqslant \prod_{k=1}^{L}\Big(1 + \frac{1}{L}R_{\mathcal{W}}\Big)\\
&\leqslant \exp(R_{\mathcal{W}})
\end{aligned}$$

On the other hand, for any $k \in \{1,\dots,L\}$,

$$\|W_k^T\|_{2,1} \leqslant \|W_k^T\|_{1,1} \leqslant R_{\mathcal{W}},$$

while

$$\|I + \frac{1}{L}W_k\| \geqslant 1 - \frac{1}{L}\|W_k\| \geqslant 1 - \frac{R_{\mathcal{W}}}{L} \geqslant \frac{1}{2},$$

under the assumption that $L \geqslant R_{\mathcal{W}}$. All in all, we obtain that

$$A(\mathbf{W}) \leqslant \exp(R_{\mathcal{W}})\big(2^{2/3}L^{1/3}R_{\mathcal{W}}^{2/3}\big)^{3/2} = 2R_{\mathcal{W}}\exp(R_{\mathcal{W}})\sqrt{L},$$

which yields the result.

B Experimental details

Our code is available at **[XXX]**.

We use the following model, corresponding to model (10) with additional projections at the beginning and at the end:

$$\begin{aligned}
H_0 &= Ax\\
H_{k+1} &= H_k + \frac{1}{L}W_{k+1}\sigma(H_k), \quad 0 \leqslant k \leqslant L-1\\
F_{\mathbf{W}}(x) &= BH_L,
\end{aligned}$$

Name	Value
d	30
L	1000
σ	ReLU

Table 1: Values of the model hyper-parameters.

where $x \in \mathbb{R}^{768}$ is a vectorized MNIST image, $A \in \mathbb{R}^{d\times 768}$, and $B \in \mathbb{R}^{10\times d}$. Table 1 gives the value of the hyper-parameters.

We use the initialization scheme outlined in Section 4.1: we initialize, for $k \in \{1, \dots, L\}$ and $i, j \in \{1, \dots, d\}$,

$$\mathbf{W}_{k,i,j} = \frac{1}{\sqrt{d}} f_{i,j}\Big(\frac{k}{L}\Big),$$

where $f_{i,j}$ are independent Gaussian processes with a RBF kernel (with bandwidth equal to 0.1). We refer to Marion et al. (2022) and Sander et al. (2022) for further discussion on this initialization scheme. However, A and B are initialized with a more usual scheme, namely with i.i.d. $\mathcal{N}(0, 1/c)$ random variables, where c denotes the number of columns of A (resp. B).

In Figure 1a, we repeat training 10 times independently. Each time, we perform 30 epochs, and compute after each epoch both the Lipschitz constant of the weights and the generalization gap. This gives 300 pairs (Lipschitz constant, generalization gap), which each corresponds to one dot in the figure. Furthermore, we report results for two setups: when A and B are trained or when they are fixed random matrices.

In Figure 1b, A and B are not trained. The reason is to assess the effect of the penalization on $\mathbf{W}$ for a fixed scale of A and B. If we allow A and B to vary, then it is possible that the effect of the penalization might be neutralized by a scale increase of A and B during training.

For all experiments, we use the standard MNIST datasplit (60k training samples and 10k testing samples). We train using the cross entropy loss, mini-batches of size 128, and the optimizer Adam (Kingma and Ba, 2015) with default parameters and a learning rate of 0.02.

We use PyTorch (Paszke et al., 2019) and PyTorch Lightning for our experiments.

The code takes about 60 hours to run on a standard laptop (no GPU).