Generalization Error Bounds for Deep Unfolding RNNs (Supplementary Material)

Boris Joukovsky [1,2] **Tanmoy Mukherjee** [1,2] **Huynh Van Luong** [1,2] **Nikos Deligiannis** [1,2]

[1]Department of Electronics and Informatics, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
[2]imec, Kapeldreef 75, B-3001 Leuven, Belgium

1 PROXIMAL OPERATOR OF REWEIGHTED-RNN

The proximal operator $\phi_{\frac{\lambda_1}{c},\frac{\lambda_2}{c},\hbar}(u)$ is defined as

$$\phi_{\frac{\lambda_1}{c},\frac{\lambda_2}{c},\hbar}(u) = \arg\min_{v\in\mathbb{R}^h} \Big\{\frac{1}{c}g(v) + \frac{1}{2}||v-u||_2^2\Big\}, \tag{1}$$

where $g(v) = \lambda_1 \mathrm{g}|v| + \lambda_2 \mathrm{g}|v-\hbar|$.

Following the proof in Appendix of [Luong et al., 2021], $\Phi_{\frac{\lambda_1}{c}\mathrm{g}_l,\frac{\lambda_2}{c}\mathrm{g}_l,\hbar}(u)$ is given by

$$\Phi_{\frac{\lambda_1}{c}\mathrm{g}_l,\frac{\lambda_2}{c}\mathrm{g}_l,\hbar}(u) = \begin{cases} u - \frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l, & \hbar + \frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l < u < \infty \\ \hbar, & \hbar + \frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l \le u \le \hbar + \frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l \\ u - \frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l, & \frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l < u < \hbar + \frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l \\ 0, & -\frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l \le u \le \frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l \\ u + \frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l, & -\infty < u < -\frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l, \end{cases} \tag{2}$$

for $\hbar \ge 0$ and

$$\Phi_{\frac{\lambda_1}{c}\mathrm{g}_l,\frac{\lambda_2}{c}\mathrm{g}_l,\hbar}(u) = \begin{cases} u - \frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l, & \frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l < u < \infty \\ 0, & -\frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l \le u \le \frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l \\ u + \frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l, & \hbar - \frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l < u < -\frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l \\ \hbar, & \hbar - \frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l \le u \le \hbar - \frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l \\ u + \frac{\lambda_1}{c}\mathrm{g}_l + \frac{\lambda_2}{c}\mathrm{g}_l, & -\infty < u < \hbar - \frac{\lambda_1}{c}\mathrm{g}_l - \frac{\lambda_2}{c}\mathrm{g}_l \end{cases} \tag{3}$$

for $\hbar < 0$.

2 SUPPORTS FOR RADEMACHER COMPLEXITY CALCULUS

The contraction lemma in Shalev-Shwartz and Ben-David [2014] shows the Rademacher complexity of the composition of a class of functions with ρ-Lipschitz functions.

Proposition 2.1. *[Shalev-Shwartz and Ben-David, 2014, Lemma 26.9—Contraction lemma]*
Let $\mathcal{F}$ be a set of functions, $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$, and $\Phi_1, ..., \Phi_m$, ρ-Lipschitz functions, namely, $|\Phi_i(\alpha) - \Phi_i(\beta)| \le \rho|\alpha - \beta|$

for all $\alpha, \beta \in \mathbb{R}$ *for some* $\rho > 0$*. For any sample set* S *of* m *points* $\mathbf{x}_1, ..., \mathbf{x}_m \in \mathcal{X}$*, let* $(\Phi \circ f)(\mathbf{x}_i) = \Phi(f(\mathbf{x}_i))$*. Then,*

$$\frac{1}{m} \underset{\boldsymbol{\epsilon} \in \{\pm 1\}^m}{\mathbb{E}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \epsilon_i (\Phi \circ f)(\mathbf{x}_i) \right] \leq \frac{\rho}{m} \underset{\boldsymbol{\epsilon} \in \{\pm 1\}^m}{\mathbb{E}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \epsilon_i f(\mathbf{x}_i) \right], \tag{4}$$

alternatively, $\mathfrak{R}_S(\boldsymbol{\Phi} \circ \mathcal{F}) \leq \rho \mathfrak{R}_S(\mathcal{F})$*, where* $\boldsymbol{\Phi}$ *denotes* $\Phi_1(\mathbf{x}_1), ..., \Phi_m(\mathbf{x}_m)$ *for* S*.*

Proposition 2.2. *[Mohri et al., 2018, Proposition A.1—Hölder's inequality]*
Let $p, q \geq 1$ *be conjugate:* $\frac{1}{p} + \frac{1}{q} = 1$*. Then, for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$*,*

$$\|\mathbf{x} \cdot \mathbf{y}\|_1 \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q, \tag{5}$$

with the equality when $|\mathrm{y}_i| = |\mathrm{x}_i|^{p-1}$ *for all* $i \in [1, n]$*.*

3 PROOF FOR GENERALIZATION ERROR BOUNDS FOR DEEP UNFOLDING RNNS

Proof. We consider the real-valued family of functions $\mathcal{F}_{d,T} : \mathbb{R}^h \times \mathbb{R}^n \mapsto \mathbb{R}$ for the functions $f^{(d)}_{\mathcal{W},\mathbf{U}}$ to update $\mathbf{h}^{(d)}_T$ in layer d, time step T, defined as

$$\mathcal{F}_{d,T} = \Big\{ (\mathbf{h}^{(d)}_{T-1}, \mathbf{x}_T) \mapsto \Phi(\mathbf{w}_d^{\mathrm{T}} f^{(d-1)}_{\mathcal{W},\mathcal{U}}(\mathbf{h}^{(d)}_{T-1}, \mathbf{x}_T) + \mathbf{u}_d^{\mathrm{T}} \mathbf{x}_T) : \|\mathbf{W}_d\|_{1,\infty} \leq \alpha_d,\ \|\mathbf{U}_d\|_{1,\infty} \leq \beta_d \Big\}, \tag{6}$$

where $\mathbf{w}_d, \mathbf{u}_d$ are the corresponding rows from $\mathbf{W}_d, \mathbf{U}_d$, respectively, and α_l, β_l, with $1 < l \leq d$, are nonnegative hyper-parameters. For the first layer and the first time step, i.e., $l = 1$, $t = 1$, the real-valued family of functions, $\mathcal{F}_{1,1} : \mathbb{R}^h \times \mathbb{R}^n \mapsto \mathbb{R}$, for the functions $f^{(1)}_{\mathcal{W},\mathcal{U}}$ is defined by:

$$\mathcal{F}_{1,1} = \Big\{ (\mathbf{h}_0, \mathbf{x}_1) \mapsto \Phi(\mathbf{w}_1^{\mathrm{T}} \mathbf{h}_0 + \mathbf{u}_1^{\mathrm{T}} \mathbf{x}_1) : \|\mathbf{W}_1\|_{1,\infty} \leq \alpha_1,\ \|\mathbf{U}\|_{1,\infty} \leq \beta_1 \Big\}, \tag{7}$$

where α_1, β_1 are nonnegative hyper-parameters. We denote the input layer as $f^{(0)}_{\mathcal{W},\mathcal{U}} = \mathbf{h}_0$ at the first time step. From the definition of Rademacher complexity [Definition 3.1] and the family of functions in (6) and (7), we obtain:

$$\begin{aligned}
m\mathfrak{R}_S(\mathcal{F}_{d,T}) &\leq \underset{\boldsymbol{\epsilon} \in \{\pm 1\}^m}{\mathbb{E}} \left[\sup_{\substack{\mathcal{W},\mathcal{U} \\ \|\mathbf{w}_d\|_1 \leq \alpha_d \\ \|\mathbf{u}_d\|_1 \leq \beta_d}} \sum_{i=1}^{m} \epsilon_i \Phi\Big(\mathbf{w}_d^{\mathrm{T}} f^{(d-1)}_{\mathcal{W},\mathcal{U}}(\mathbf{h}_{T-1,i}, \mathbf{x}_{T,i}) + \mathbf{u}_d^{\mathrm{T}} \mathbf{x}_{T,i} \Big) \right] \\
&\leq \frac{1}{\lambda} \log \exp \left(\underset{\boldsymbol{\epsilon} \in \{\pm 1\}^m}{\mathbb{E}} \left[\sup_{\substack{\mathcal{W},\mathcal{U} \\ \|\mathbf{w}_d\|_1 \leq \alpha_d \\ \|\mathbf{u}_d\|_1 \leq \beta_d}} \lambda \sum_{i=1}^{m} \epsilon_i \Big(\mathbf{w}_d^{\mathrm{T}} f^{(d-1)}_{\mathcal{W},\mathcal{U}}(\mathbf{h}_{T-1,i}, \mathbf{x}_{T,i}) + \mathbf{u}_d^{\mathrm{T}} \mathbf{x}_{T,i} \Big) \right] \right) \\
&\leq \frac{1}{\lambda} \log \underset{\boldsymbol{\epsilon} \in \{\pm 1\}^m}{\mathbb{E}} \left[\sup_{\substack{\mathcal{W},\mathcal{U} \\ \|\mathbf{w}_d\|_1 \leq \alpha_d \\ \|\mathbf{u}_d\|_1 \leq \beta_d}} \exp \left(\lambda \sum_{i=1}^{m} \epsilon_i \Big(\mathbf{w}_d^{\mathrm{T}} f^{(d-1)}_{\mathcal{W},\mathcal{U}}(\mathbf{h}_{T-1,i}, \mathbf{x}_{T,i}) \Big) + \lambda \sum_{i=1}^{m} \epsilon_i \mathbf{u}_d^{\mathrm{T}} \mathbf{x}_{T,i} \right) \right] \qquad \text{(8a)} \\
&\leq \frac{1}{\lambda} \log \underset{\boldsymbol{\epsilon} \in \{\pm 1\}^m}{\mathbb{E}} \left[\sup_{\substack{\mathcal{W},\mathcal{U} \\ \|\mathbf{w}_d\|_1 \leq \alpha_d}} \exp \left(\lambda \sum_{i=1}^{m} \epsilon_i \Big(\mathbf{w}_d^{\mathrm{T}} f^{(d-1)}_{\mathcal{W},\mathcal{U}}(\mathbf{h}_{T-1,i}, \mathbf{x}_{T,i}) \Big) \right) \sup_{\|\mathbf{u}_d\|_1 \leq \beta_d} \exp \left(\lambda \sum_{i=1}^{m} \epsilon_i \mathbf{u}_d^{\mathrm{T}} \mathbf{x}_{T,i} \right) \right], \qquad \text{(8b)}
\end{aligned}$$

where $\lambda > 0$ is an arbitrary parameter, Eq. (8a) follows Lemma 2.1 for 1-Lipschitz Φ a long with Inequality (20), and (8b) holds by Inequality (17).

For layer $1 \leq l \leq d$ and time step t, let us denote:

$$\Delta^{(l)}_{\mathbf{h}_{t-1},\mathbf{x}_t} = \sup_{\substack{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}} \\ \|\mathbf{w}_l\|_1 \leq \alpha_l}} \exp\left(\lambda \Lambda_l \sum_{i=1}^{m} \epsilon_i \Big(\mathbf{w}_l^{\mathrm{T}} f^{(l-1)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{t-1,i}, \mathbf{x}_{t,i})\Big)\right), \tag{9}$$

$$\Delta^{(l)}_{\mathbf{x}_t} = \sup_{\|\mathbf{u}_l\|_1 \leq \beta_l} \exp\left(\lambda \Lambda_l \sum_{i=1}^{m} \epsilon_i \Big(\mathbf{u}_l^{\mathrm{T}} \mathbf{x}_{t,i}\Big)\right), \tag{10}$$

where Λ_l is defined as follows: $\Lambda_d = 1$, $\Lambda_l = \prod_{k=l+1}^{d} \alpha_k$ with $1 \leq l \leq d-1$, and $\Lambda_0 = \prod_{k=1}^{d} \alpha_k$.

Following the Hölder's inequality in (5) in case of $p = 1$ and $q = \infty$ applied to $\mathbf{w}_l^{\mathrm{T}}$ and $f^{(l-1)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{t-1,i}, \mathbf{x}_{t,i})$ in (9), respectively, we get:

$$\begin{aligned}
\Delta^{(d)}_{\mathbf{h}_{t-1},\mathbf{x}_t} &\leq \sup_{\substack{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}} \\ \|\mathbf{W}_{d-1}\|_{1,\infty} \leq \alpha_{d-1} \\ \|\mathbf{U}_{d-1}\|_{1,\infty} \leq \beta_{d-1}}} \exp\left(\lambda \alpha_d \left\| \sum_{i=1}^{m} \epsilon_i \Phi\Big(\mathbf{W}_{d-1} f^{(d-2)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{t-1,i}, \mathbf{x}_{t,i}) + \mathbf{U}_{d-1}\mathbf{x}_{t,i}\Big)\right\|_{\infty}\right) \\
&\leq \sup_{\substack{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}} \\ \|\mathbf{w}_{d-1,k}\|_1 \leq \alpha_{d-1} \\ \|\mathbf{u}_{d-1,k}\|_1 \leq \beta_{d-1}}} \exp\left(\lambda \alpha_d \max_{k \in \{1,\cdots,h\}} \left| \sum_{i=1}^{m} \epsilon_i \Phi\Big(\mathbf{w}_{d-1,k}^{\mathrm{T}} f^{(d-2)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{t-1,i}, \mathbf{x}_{t,i}) + \mathbf{u}_{d-1,k}^{\mathrm{T}}\mathbf{x}_{t,i}\Big)\right|\right) \\
&\leq \sup_{\substack{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}} \\ \|\mathbf{w}_{d-1,k}\|_1 \leq \alpha_{d-1} \\ \|\mathbf{u}_{d-1,k}\|_1 \leq \beta_{d-1}}} \exp\left(\lambda \alpha_d \left| \sum_{i=1}^{m} \epsilon_i \Phi\Big(\mathbf{w}_{d-1,k}^{\mathrm{T}} f^{(d-2)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{t-1,i}, \mathbf{x}_{t,i}) + \mathbf{u}_{d-1,k}^{\mathrm{T}}\mathbf{x}_{t,i}\Big)\right|\right).
\end{aligned} \tag{11}$$

Similarly, from (10), we obtain:

$$\Delta^{(d)}_{\mathbf{x}_t} \leq \sup_{\|\mathbf{u}_d\|_1 \leq \beta_d} \exp\left(\lambda \sum_{i=1}^{m} \epsilon_i \mathbf{u}_d^{\mathrm{T}} \mathbf{x}_{t,i}\right) \leq \exp\left(\lambda \beta_d \left\| \sum_{i=1}^{m} \epsilon_i \mathbf{x}_{t,i} \right\|_{\infty}\right) \leq \exp\left(\lambda \beta_d \left| \sum_{i=1}^{m} \epsilon_i \mathrm{x}_{\tau,i,\kappa} \right|\right), \tag{12}$$

where $\{\tau, \kappa\} = \underset{t \in \{1,\ldots,T\}, j \in \{1,\ldots,n\}}{\operatorname{argmax}} \left| \sum_{i=1}^{m} \epsilon_i \mathrm{x}_{t,i,j} \right|$.

From (8b), (11), and (12), we get:

$$\begin{aligned}
&m\mathfrak{R}_S(\mathcal{F}_{d,T}) \\
&\leq \frac{1}{\lambda} \log\left(\mathop{\mathbb{E}}_{\boldsymbol{\epsilon} \in \{\pm 1\}^m} \left[\sup_{\substack{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}} \\ \|\mathbf{w}_{d-1,k}\|_1 \leq \alpha_{d-1} \\ \|\mathbf{u}_{d-1,k}\|_1 \leq \beta_{d-1}}} \exp\left(\lambda \alpha_d \left| \sum_{i=1}^{m} \epsilon_i \Phi\Big(\mathbf{w}_{d-1,k}^{\mathrm{T}} f^{(d-2)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{T-1,i}, \mathbf{x}_{T,i}) + \mathbf{u}_{d-1,k}^{\mathrm{T}}\mathbf{x}_{T,i}\Big)\right| + \lambda \beta_d \left| \sum_{i=1}^{m} \epsilon_i \mathrm{x}_{\tau,i,\kappa} \right|\right)\right]\right) \\
&\leq \frac{1}{\lambda} \log\left(\mathop{\mathbb{E}}_{\boldsymbol{\epsilon} \in \{\pm 1\}^m} \left[\sup_{\substack{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}} \\ \|\mathbf{w}_{d-1,k}\|_1 \leq \alpha_{d-1} \\ \|\mathbf{u}_{d-1,k}\|_1 \leq \beta_{d-1}}} \left(\exp\left(\lambda \alpha_d \sum_{i=1}^{m} \epsilon_i \Phi\Big(\mathbf{w}_{d-1,k}^{\mathrm{T}} f^{(d-2)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{T-1,i}, \mathbf{x}_{T,i}) + \mathbf{u}_{d-1,k}^{\mathrm{T}}\mathbf{x}_{T,i}\Big) + \lambda \beta_d \sum_{i=1}^{m} \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right) \right.\right.\right. \\
&\quad + \exp\left(\lambda \alpha_d \sum_{i=1}^{m} \epsilon_i \Phi\Big(\mathbf{w}_{d-1,k}^{\mathrm{T}} f^{(d-2)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{T-1,i}, \mathbf{x}_{T,i}) + \mathbf{u}_{d-1,k}^{\mathrm{T}}\mathbf{x}_{T,i}\Big) - \lambda \beta_d \sum_{i=1}^{m} \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right) \\
&\quad + \exp\left(-\lambda \alpha_d \sum_{i=1}^{m} \epsilon_i \Phi\Big(\mathbf{w}_{d-1,k}^{\mathrm{T}} f^{(d-2)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{T-1,i}, \mathbf{x}_{T,i}) + \mathbf{u}_{d-1,k}^{\mathrm{T}}\mathbf{x}_{T,i}\Big) + \lambda \beta_d \sum_{i=1}^{m} \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right) \\
&\quad \left.\left.\left. + \exp\left(-\lambda \alpha_d \sum_{i=1}^{m} \epsilon_i \Phi\Big(\mathbf{w}_{d-1,k}^{\mathrm{T}} f^{(d-2)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{T-1,i}, \mathbf{x}_{T,i}) + \mathbf{u}_{d-1,k}^{\mathrm{T}}\mathbf{x}_{T,i}\Big) - \lambda \beta_d \sum_{i=1}^{m} \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right)\right)\right]\right)
\end{aligned}$$

$$\leq \frac{1}{\lambda}\log\left(4\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\Delta^{(d-1)}_{\mathbf{h}_{T-1},\mathbf{x}_T}\Delta^{(d-1)}_{\mathbf{x}_T}\exp\left(\beta_d\lambda\sum_{i=1}^m \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right)\right]\right) \tag{13a}$$

$$\leq \frac{1}{\lambda}\log\left(4^{d-1}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\Delta^{(1)}_{\mathbf{h}_{T-1},\mathbf{x}_T}\Delta^{(1)}_{\mathbf{x}_T}\exp\left(\lambda\Big(\sum_{l=2}^d \beta_l\Lambda_l\Big)\sum_{i=1}^m \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right)\right]\right) \tag{13b}$$

$$\leq \frac{1}{\lambda}\log\left(4^{d-1}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(\lambda\Big(\sum_{l=2}^d \beta_l\Lambda_l\Big)\sum_{i=1}^m \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right)\sup_{\|\mathbf{w}_1\|_1\leq\alpha_1}\exp\left(\lambda\Lambda_1\sum_{i=1}^m\epsilon_i\Big(\mathbf{w}_1^{\mathrm{T}}\mathbf{h}_{T-1,i}\Big)\right)\right.\right.$$
$$\left.\left.\cdot\sup_{\|\mathbf{u}_1\|_1\leq\beta_1}\exp\left(\lambda\Lambda_1\sum_{i=1}^m\epsilon_i\Big(\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_{T,i}\Big)\right)\right]\right) \tag{13c}$$

$$\leq \frac{1}{\lambda}\log\left(4^{d-1}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(\lambda\Big(\sum_{l=2}^d \beta_l\Lambda_l\Big)\sum_{i=1}^m \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right)\sup_{\substack{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}\\ \|\mathbf{w}_d\|_1\leq\alpha_d\\ \|\mathbf{u}_d\|_1\leq\beta_d}}\exp\left(\lambda\Lambda_0\Big\|\sum_{i=1}^m\epsilon_i\mathbf{h}_{T-1,i}\Big\|_\infty\right)\right.\right.$$
$$\left.\left.\cdot\exp\left(\lambda\beta_1\Lambda_1\Big\|\sum_{i=1}^m\epsilon_i\mathbf{x}_{T,i}\Big\|_\infty\right)\right]\right) \tag{13d}$$

$$\leq \frac{1}{\lambda}\log\left(4^{d}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(\lambda\Big(\sum_{l=1}^d \beta_l\Lambda_l\Big)\sum_{i=1}^m \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right)\right.\right.$$
$$\left.\left.\cdot\sup_{\substack{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}\\ \|\mathbf{w}_d\|_1\leq\alpha_d\\ \|\mathbf{u}_d\|_1\leq\beta_d}}\exp\left(\lambda\Lambda_0\sum_{i=1}^m\epsilon_i\Phi\Big(\mathbf{w}_d^{\mathrm{T}}f^{(d-1)}_{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{U}}}(\mathbf{h}_{T-2,i},\mathbf{x}_{T-1,i})+\mathbf{u}_d^{\mathrm{T}}\mathbf{x}_{T-1,i}\Big)\right)\right]\right), \tag{13e}$$

where (13a) holds by Inequality (17) and (13b) follows by repeating the process from layer $d-1$ to layer 1 for time step T. Furthermore, (13c) is returned as the beginning of the process for time step $T-1$ and (13d) follows Inequality (5).

Proceeding by repeating the above procedure in (13e) from time step $T-1$ to time step 1, we get:

$$m\mathfrak{R}_S(\mathcal{F}_{d,T})\leq \frac{1}{\lambda}\log\left(4^{dT}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(\lambda\Big(\sum_{l=1}^d \beta_l\Lambda_l\Big)\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)\sum_{i=1}^m \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right)\exp\left(\lambda\Lambda_0^T\Big\|\sum_{i=1}^m\epsilon_i\mathbf{h}_0\Big\|_\infty\right)\right]\right). \tag{14}$$

Let us denote $\mu=\mathop{\mathrm{argmax}}_{j\in\{1,\ldots,h\}}\Big|\sum_{i=1}^m\epsilon_i\mathrm{h}_{0,j}\Big|$, from (14), we have:

$$m\mathfrak{R}_S(\mathcal{F}_{d,T})\leq \frac{1}{\lambda}\log\left(4^{dT}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(\lambda\Big(\sum_{l=1}^d \beta_l\Lambda_l\Big)\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)\sum_{i=1}^m \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right)\exp\left(\lambda\Lambda_0^T\sum_{i=1}^m\epsilon_i\mathrm{h}_{0,\mu}\right)\right]\right)$$
$$\leq\frac{2dT\log2}{\lambda}+\frac{1}{2\lambda}\log\left(\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(\lambda\Big(\sum_{l=1}^d \beta_l\Lambda_l\Big)\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)\sum_{i=1}^m \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right)\exp\left(\lambda\Lambda_0^T\sum_{i=1}^m\epsilon_i\mathrm{h}_{0,\mu}\right)\right]\right)^2$$
$$\leq\frac{2dT\log2}{\lambda}+\frac{1}{2\lambda}\log\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(2\lambda\Big(\sum_{l=1}^d \beta_l\Lambda_l\Big)\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)\sum_{i=1}^m \epsilon_i \mathrm{x}_{\tau,i,\kappa}\right)\right]+\frac{1}{2\lambda}\log\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(2\lambda\Lambda_0^T\sum_{i=1}^m\epsilon_i\mathrm{h}_{0,\mu}\right)\right] \tag{15a}$$

$$\leq\frac{2dT\log2}{\lambda}+\frac{1}{2\lambda}\log\sum_{j=1}^n\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(2\lambda\Big(\sum_{l=1}^d \beta_l\Lambda_l\Big)\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)\sum_{i=1}^m \epsilon_i \mathrm{x}_{\tau,i,j}\right)\right]$$
$$+\frac{1}{2\lambda}\log\sum_{j=1}^h\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(2\lambda\Lambda_0^T\sum_{i=1}^m\epsilon_i\mathrm{h}_{0,j}\right)\right] \tag{15b}$$

$$\leq\frac{2dT\log2}{\lambda}+\frac{1}{2\lambda}\log\sum_{j=1}^n\prod_{i=1}^m\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\in\{\pm1\}^m}\left[\exp\left(2\lambda\Big(\sum_{l=1}^d \beta_l\Lambda_l\Big)\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)\epsilon_i \mathrm{x}_{\tau,i,j}\right)\right]$$

$$+\frac{1}{2\lambda}\log\sum_{j=1}^{h}\prod_{i=1}^{m}\mathop{\mathbb{E}}_{\epsilon\in\{\pm1\}^m}\left[\exp\left(2\lambda\Lambda_0^T\epsilon_i \mathrm{h}_{0,j}\right)\right]$$

$$\leq\frac{2dT\log 2}{\lambda}+\frac{1}{2\lambda}\log\sum_{j=1}^{n}\prod_{i=1}^{m}\left[\frac{1}{2}\exp\left(2\lambda\Big(\sum_{l=1}^{d}\beta_l\Lambda_l\Big)\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)\mathrm{x}_{\tau,i,j}\right)+\frac{1}{2}\exp\left(-2\lambda\Big(\sum_{l=1}^{d}\beta_l\Lambda_l\Big)\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)\mathrm{x}_{\tau,i,j}\right)\right]$$

$$+\frac{1}{2\lambda}\log\sum_{j=1}^{h}\prod_{i=1}^{m}\left[\frac{1}{2}\exp\left(2\lambda\Lambda_0^T\mathrm{h}_{0,j}\right)+\frac{1}{2}\exp\left(-2\lambda\Lambda_0^T\mathrm{h}_{0,j}\right)\right]$$

$$\leq\frac{2dT\log 2}{\lambda}+\frac{1}{2\lambda}\log\sum_{j=1}^{n}\left[\exp\left(2\lambda^2\Big(\sum_{l=1}^{d}\beta_l\Lambda_l\Big)^2\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)^2\sum_{i=1}^{m}x_{\tau,i,j}^2\right)\right]+\frac{1}{2\lambda}\log\sum_{j=1}^{h}\left[\exp\left(2\lambda^2\Lambda_0^{2T}\sum_{i=1}^{m}h_{0,j}^2\right)\right] \tag{15c}$$

$$\leq\frac{2dT\log 2}{\lambda}+\frac{\log n}{2\lambda}+\lambda\Big(\sum_{l=1}^{d}\beta_l\Lambda_l\Big)^2\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)^2 mB_{\mathbf{x}}^2+\frac{\log h}{2\lambda}+\lambda\Lambda_0^{2T}m\|\mathbf{h}_0\|_\infty^2$$

$$\leq\frac{2dT\log 2+\log\sqrt{n}+\log\sqrt{h}}{\lambda}+\lambda\left(\Big(\sum_{l=1}^{d}\beta_l\Lambda_l\Big)^2\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)^2 mB_{\mathbf{x}}^2+\Lambda_0^{2T}m\|\mathbf{h}_0\|_\infty^2\right), \tag{15d}$$

where (15a) follows Inequality (19), (15b) holds by replacing with $\sum_{j=1}^{n}$ and $\sum_{j=1}^{h}$, respectively. In addition, (15c) follows (18) and (15d) is received by the following definition: At time step t, we define $\mathbf{X}_t\in\mathbb{R}^{n\times m}$, a matrix composed of m columns from the m input vectors $\{\mathbf{x}_{t,i}\}_{i=1}^{m}$; we also define $\|\mathbf{X}_t\|_{2,\infty}=\sqrt{\max_{k\in\{1,\dots,n\}}\sum_{i=1}^{m}\mathrm{x}_{t,i,k}^2}\leq\sqrt{m}B_{\mathbf{x}}$, representing the maximum of the ℓ_2-norms of the rows of matrix $\mathbf{X}_t$, and $\|\mathbf{h}_0\|_\infty=\max_j|\mathrm{h}_{0,j}|$.

Choosing $\lambda=\sqrt{\frac{2dT\log 2+\log\sqrt{n}+\log\sqrt{h}}{\Big(\sum_{l=1}^{d}\beta_l\Lambda_l\Big)^2\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)^2 mB_{\mathbf{x}}^2+\Lambda_0^{2T}m\|\mathbf{h}_0\|_\infty^2}}$, we achieve the upper bound:

$$\mathfrak{R}_S(\mathcal{F}_{d,T})\leq\sqrt{\frac{2(4dT\log 2+\log n+\log h)}{m}\left(\Big(\sum_{l=1}^{d}\beta_l\Lambda_l\Big)^2\Big(\frac{\Lambda_0^T-1}{\Lambda_0-1}\Big)^2 B_{\mathbf{x}}^2+\Lambda_0^{2T}\|\mathbf{h}_0\|_\infty^2\right)}. \tag{16}$$

It can be noted that $\mathfrak{R}_S(\mathcal{F}_{d,T})$ in (16) is derived for the real-valued functions $\mathcal{F}_{d,T}$. For the vector-valued functions $\mathcal{F}_{d,T}:\mathbb{R}^h\times\mathbb{R}^n\mapsto\mathbb{R}^h$ [in Theorem 3.3]we apply the contraction lemma [Lemma 2.1] to a Lipschitz loss to obtain the complexity of such vector-valued functions by means of the complexity of the real-valued functions. Specifically, in Theorem 3.3 under the assumption of the 1-Lipschitz loss function and from Theorem 3.2, Lemma 2.1, we complete the proof.

□

Supporting inequalities:

(i) If A, B are sets of positive real numbers, then:

$$\sup(AB)=\sup(A)\cdot\sup(B). \tag{17}$$

(ii) Given $x\in\mathbb{R}$, we have:

$$\frac{\exp(x)+\exp(-x)}{2}\leq\exp(x^2/2). \tag{18}$$

(iii) Let X and Y be random variables, the Cauchy–Bunyakovsky–Schwarz inequality gives:

$$(\mathbb{E}[XY])^2\leq\mathbb{E}[X^2]\cdot\mathbb{E}[Y^2]. \tag{19}$$

(iv) If ψ is a convex function, the Jensen's inequality gives:

$$\psi(\mathbb{E}[X])\leq\mathbb{E}[\psi(X)]. \tag{20}$$

4 EXTENSION OF GENERALIZATION ERROR BOUND FOR CLASSIFICATION

Proof. Let $\mathbf{y} = \mathbf{Y}\mathbf{h}_t^{(d)} \equiv y(\mathbf{h}_t^{(d)})$ be a linear classifier with $\mathbf{Y} \in \mathbb{R}^{c\times h}$. Let $\mathbf{Y}_i$ denote the i^{th} row of $\mathbf{Y}$. Below, we show that each entry $y_i(\mathbf{h}_t^{(d)})$ is ρ-Lipschitz on its input with $\rho = \min(\max_i \|\mathbf{Y}_i\|_2, \max_i \|\mathbf{Y}_i\|_1)$:

$$\forall \mathbf{h}, \mathbf{h}' \in \mathbb{R}^h, \forall i \in \{1, ..., c\}: \quad |y_i - y_i'| = \left|\mathbf{Y}_i^T\mathbf{h} - \mathbf{Y}_i^T\mathbf{h}'\right| = \left|\mathbf{Y}_i^T(\mathbf{h} - \mathbf{h}')\right| \tag{21}$$
$$\leq \|\mathbf{Y}_i\|_2 \left\|\mathbf{h} - \mathbf{h}'\right\|_2 \tag{22}$$
$$\leq \max_{j\in\{1,\dots,c\}} \|\mathbf{Y}_j\|_2 \left\|\mathbf{h} - \mathbf{h}'\right\|_2 \tag{23}$$

In the development above, line 22 was obtained by applying the triangular inequality. Moreover, in line 23, we have identified a unique Lipschitz constant that is valid for all i. Alternatively, we can also write that:

$$\forall \mathbf{h}, \mathbf{h}' \in \mathbb{R}^h, \forall i \in \{1, ..., c\}: \quad |y_i - y_i'| \leq \|\mathbf{Y}_i\|_1 \left\|\mathbf{h} - \mathbf{h}'\right\|_\infty \tag{24}$$
$$\leq \max_{j\in\{1,\dots,c\}} \|\mathbf{Y}_j\|_1 \left\|\mathbf{h} - \mathbf{h}'\right\|_\infty \tag{25}$$
$$\leq \max_{j\in\{1,\dots,c\}} \|\mathbf{Y}_j\|_1 \left\|\mathbf{h} - \mathbf{h}'\right\|_2 \tag{26}$$

In the development above, line 25 was obtained using Hölder's inequality [see Proposition 2.2] and line 26 was obtained considering that the ℓ_2 norm is an upper bound of the ℓ_∞ norm. Setting $\rho = \min\{\max_i \|\mathbf{Y}_i\|_2, \max_i \|\mathbf{Y}_i\|_1\}$ completes the proof and shows that ρ is a Lipschitz constant for each entry $y_i(\mathbf{h}_t^{(d)})$. To obtain the generalization upper bound proposed in section 4.3 using the ramp loss evaluated on the classification margin $\ell_\gamma(\mathbf{y})$ (which is $\frac{1}{\gamma}$-Lipschitz), it suffies to apply the contraction lemma twice [see Proposition 2.1], first for the composition with multivariate linear classifier function and secondly with the ℓ_γ loss function, leading to:

$$L_{\mathcal{D},\gamma}(f) - L_{S,\gamma}(f) \leq \frac{2\rho}{\gamma}\mathfrak{R}_S(\mathcal{F}_{d,T}) + 4\sqrt{\frac{2\log(4/\delta)}{m}} \tag{27}$$

□

References

Huynh Van Luong, Boris Joukovsky, and Nikos Deligiannis. Designing interpretable recurrent neural networks for video reconstruction via deep unfolding. *IEEE Transactions on Image Processing*, 30:4099–4113, 2021.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning, Second edition*. MIT Press, Cambridge, Massachusetts, USA, 2018. ISBN 9780262039406.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.