A Novel Optimization Method for Recurrent Neural Network by Non-Euclidean Geometry - Appendix

Chen Liu
Master Student, IC, EPFL
chen.liu@epfl.ch

Ya-Ping Hsieh
Supervisor, PH.D, EL, EPFL
ya-ping.hsieh@epfl.ch

Volkan Cevher
Supervisor, Associate Professor, EL, EPFL
volkan.cevher@epfl.ch

A General Upper Bound of RNN Loss Function

Let $\mathbf{\Phi} = \{\mathbf{U}, \mathbf{W}, \mathbf{V}, \mathbf{s}_0\}$.We already have the expression of t-th token's contribution to loss function.

$$E_t(\mathbf{\Phi}) = \log \sum_{j=1}^{K} e^{\mathbf{V}_j \mathbf{s}_t} - \mathbf{y}^T \mathbf{V}\mathbf{s}_t \tag{1}$$

This function contains two parts: log-sum-exp part and linear part. Let $f_t(\mathbf{\Phi}) = \log \sum_{j=1}^{K} e^{\mathbf{V}_j \mathbf{s}_t}$, $g_t(\mathbf{\Phi}) = -\mathbf{y}^T \mathbf{V}\mathbf{s}_t$ and $\mathbf{r} = \mathbf{V}\mathbf{s}_t(\mathbf{\Phi} + \Delta\mathbf{\Phi}) - \mathbf{V}\mathbf{s}_t(\mathbf{\Phi}) - \langle \nabla_{\mathbf{\Phi}} \mathbf{V}\mathbf{s}_t(\mathbf{\Phi}), \Delta\mathbf{\Phi} \rangle$. We will obtain the upper bound of formula 1 using the following two lemmas.

Lemma 1 $f_t(\mathbf{\Phi} + \Delta\mathbf{\Phi}) \leq f_t(\mathbf{\Phi}) + \langle \nabla_{\mathbf{\Phi}} f_t(\mathbf{\Phi}), \Delta\mathbf{\Phi} \rangle + \frac{1}{2} \|\mathbf{V}\mathbf{s}_t(\mathbf{\Phi} + \Delta\mathbf{\Phi}) - \mathbf{V}\mathbf{s}_t(\mathbf{\Phi})\|_\infty^2 + \|r\|_\infty$

According to [1], for log-sum-exp function $h(\mathbf{x}) = \log \sum_{j=1}^{J} e^{\mathbf{x}_j}$, Lipschitz constant $L_\infty = \frac{1}{2}$ and we can derive an upper bound of function $h(\mathbf{x})$.

$$h(\mathbf{x} + \Delta\mathbf{x}) \leq h(\mathbf{x}) + \langle \nabla_{\mathbf{x}} h(\mathbf{x}), \Delta\mathbf{x} \rangle + \frac{1}{2} \|\Delta\mathbf{x}\|_\infty^2 \tag{2}$$

It is not difficult to find $f_t(\mathbf{\Phi}) = h(\mathbf{V}\mathbf{s}_t(\mathbf{\Phi}))$. Plug this into equation 2, we have the following formula.

$$f_t(\mathbf{\Phi} + \Delta\mathbf{\Phi}) \leq f(\mathbf{\Phi}) + \langle \nabla_{\mathbf{V}\mathbf{s}_t} f_t(\mathbf{\Phi}), \mathbf{V}\mathbf{s}_t(\mathbf{\Phi} + \Delta\mathbf{\Phi}) - \mathbf{V}\mathbf{s}_t(\mathbf{\Phi}) \rangle + \frac{1}{2} \|\mathbf{V}\mathbf{s}_t(\mathbf{\Phi} + \Delta\mathbf{\Phi}) - \mathbf{V}\mathbf{s}_t(\mathbf{\Phi})\|_\infty \tag{3}$$

From the definition of $\mathbf{r}$, we have $\mathbf{V}\mathbf{s}_t(\mathbf{\Phi} + \Delta\mathbf{\Phi}) - \mathbf{V}\mathbf{s}_t(\mathbf{\Phi}) = \mathbf{r} + \langle \nabla_{\mathbf{\Phi}} \mathbf{V}\mathbf{s}_t(\mathbf{\Phi}), \Delta\mathbf{\Phi} \rangle$. Subsequently, we have the following equation.

$$\begin{aligned} &\langle \nabla_{\mathbf{V}\mathbf{s}_t} f_t(\mathbf{\Phi}), \mathbf{V}\mathbf{s}_t(\mathbf{\Phi} + \Delta\mathbf{\Phi}) - \mathbf{V}\mathbf{s}_t(\mathbf{\Phi}) \rangle = \langle \nabla_{\mathbf{V}\mathbf{s}_t} f_t(\mathbf{\Phi}), \mathbf{r} + \langle \nabla_{\mathbf{\Phi}} \mathbf{V}\mathbf{s}_t(\mathbf{\Phi}), \Delta\mathbf{\Phi} \rangle \rangle \\ &= \mathbf{r}^T \nabla_{\mathbf{V}\mathbf{s}_t} f_t(\mathbf{\Phi}) + \langle \nabla_{\mathbf{V}\mathbf{s}_t} f_t(\mathbf{\Phi}) \nabla_{\mathbf{\Phi}} \mathbf{V}\mathbf{s}_t(\mathbf{\Phi}), \Delta\mathbf{\Phi} \rangle = \mathbf{r}^T \nabla_{\mathbf{V}\mathbf{s}_t} f_t(\mathbf{\Phi}) + \langle \nabla_{\mathbf{\Phi}} f_t(\mathbf{\Phi}), \Delta\mathbf{\Phi} \rangle \end{aligned} \tag{4}$$

From $\nabla_{\mathbf{V}\mathbf{s}_t} f_t(\mathbf{\Phi}) = softmax(\mathbf{V}\mathbf{s}_t)$, we have $\mathbf{r}^T \nabla_{\mathbf{V}\mathbf{s}_t} f_t(\mathbf{\Phi}) \leq \|\mathbf{r}\|_\infty \|softmax(\mathbf{V}\mathbf{s}_t)\|_\infty \leq \|\mathbf{r}\|_\infty$. Combine this with equation 3 and 4, we will get **Lemma 1**. #

Lemma 2 $g_t(\mathbf{\Phi} + \mathbf{\Delta\Phi}) \leq \langle \nabla_{\mathbf{\Phi}} g_t(\mathbf{\Phi}), \Delta\mathbf{\Phi} \rangle + \|\mathbf{r}\|_\infty$

$g_t(\mathbf{\Phi})$ is linear, so it is equal to its first-order Taylor expansion.

$$g_t(\mathbf{\Phi} + \Delta\mathbf{\Phi}) = g_t(\mathbf{\Phi}) + \langle \nabla_{\mathbf{V}\mathbf{s}_t} g_t(\mathbf{\Phi}), \mathbf{V}\mathbf{s}_t(\mathbf{\Phi} + \Delta\mathbf{\Phi}) - \mathbf{V}\mathbf{s}_t(\mathbf{\Phi}) \rangle \tag{5}$$

Similar to formula 4, we have the following upper bound expression.

$$\langle \nabla_{\mathbf{Vs}_t} g_t(\mathbf{\Phi}), \mathbf{Vs}_t(\mathbf{\Phi}+\Delta\mathbf{\Phi}) - \mathbf{Vs}_t(\mathbf{\Phi}) \rangle \leq \langle \nabla_{\mathbf{\Phi}} g_t(\mathbf{\Phi}), \Delta\mathbf{\Phi} \rangle + \|\mathbf{r}\|_\infty \tag{6}$$

Combine formula 5 with 6, we will get **Lemma 2** immediately. #

The general upper bound of RNN loss function below is obtained by adding **Lemma 1** and **Lemma 2**.

$$\begin{aligned} E_t(\mathbf{\Phi}+\Delta\mathbf{\Phi}) \leq E_t(\mathbf{\Phi}) + \langle \nabla_{\mathbf{\Phi}} E_t(\mathbf{\Phi}), \Delta\mathbf{\Phi} \rangle + \frac{1}{2}\|\mathbf{Vs}_t(\mathbf{\Phi}+\Delta\mathbf{\Phi}) - \mathbf{Vs}_t(\mathbf{\Phi})\|_\infty^2 + \\ 2\|\mathbf{Vs}_t(\mathbf{\Phi}+\Delta\mathbf{\Phi}) - \mathbf{Vs}_t(\mathbf{\Phi}) - \langle \mathbf{V}\nabla\mathbf{s}_t(\mathbf{\Phi}), \Delta\mathbf{\Phi} \rangle\|_\infty \end{aligned} \tag{7}$$

B Lipschitz Constants of Some Matrices

In this section, we will derive the upper bound of Lipschitz constants of RNN's input matrix and recurrent matrix. We will repeatedly use the following inequality in the following subsections. The proof of them are available in the section 9.3 in [2] or in the appendix of [3].

$$if\ q^{-1}+p^{-1}=r^{-1}\ and\ p,q\in[1,+\infty],\ then\ \|\mathbf{M}\|_{S^r} \leq \|\mathbf{M}\|_{S^q}\|\mathbf{M}\|_{S^p} \tag{8}$$

$$if\ \mathbf{v}\ is\ a\ vector,\ then\ \|\mathbf{r}\|_{S^p} = \|\mathbf{r}\|_2\ \forall p\in[1,+\infty] \tag{9}$$

$$tr(\mathbf{M}_1\mathbf{M}_2) \leq \|\mathbf{M}_1\|_{S^1}\|\mathbf{M}_2\|_{S^\infty} \tag{10}$$

$$\|\mathbf{x}\odot\mathbf{y}\|_2 \leq \|\mathbf{x}\|_\infty\|\mathbf{y}\|_2 \tag{11}$$

Here, $\|\mathbf{M}\|_{S^p}$ means Schatten-p norm of a matrix, $tr(\mathbf{M})$ is the trace of a matrix and $\odot$ means element-wise multiplication. The notation in this section is the same as the main body.

B.1 Lipschitz Constant of the Input Matrix

According to inequality 7, we have two parts to be upper bounded: $\|\mathbf{Vs}_t(\mathbf{U}^k+\Delta\mathbf{U}) - \mathbf{Vs}_t(\mathbf{U}^k)\|_\infty$ and $\|\mathbf{Vs}_t(\mathbf{U}^k+\Delta\mathbf{U}) - \mathbf{Vs}_t(\mathbf{U}^k) - \langle \mathbf{V}\nabla\mathbf{s}_t(\mathbf{U}^k), \Delta\mathbf{U}\rangle\|_\infty$.

The first part is easy. We can approximate $\mathbf{s}_t(\mathbf{U}+\Delta\mathbf{U})$ using its first order Taylor expansion.

$$[\mathbf{Vs}_t(\mathbf{U}^k+\Delta\mathbf{U}) - \mathbf{Vs}_t(\mathbf{U}^k)]_p \simeq tr[\frac{\partial[\mathbf{Vs}_t]_p}{\partial\mathbf{U}}\Delta\mathbf{U}] \leq \|\mathbf{Q}_t\tilde{\mathbf{V}}_p\|_{S^1}\|\Delta\mathbf{U}\|_{S^\infty} \tag{12}$$

The last less than sign satisfies using formula 10 and $\frac{\partial\mathbf{s}_t(\mathbf{U})}{\partial\mathbf{U}} = \mathbf{Q}_t$. So, the upper bound of first part is:

$$\|\mathbf{Vs}_t(\mathbf{U}^k+\Delta\mathbf{U}) - \mathbf{Vs}_t(\mathbf{U}^k)\|_\infty \leq \max_p \|\mathbf{Q}_t\tilde{\mathbf{V}}_p\|_{S^1}\|\Delta\mathbf{U}\|_{S^\infty} \tag{13}$$

For the second part, we use Taylor expansion to analyze its p-th element.

$$\begin{aligned} &[\mathbf{Vs}_t(\mathbf{U}^k+\Delta\mathbf{U}) - \mathbf{Vs}_t(\mathbf{U}^k) - \langle \mathbf{V}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k), \Delta\mathbf{U}\rangle]_p \\ =& \int_0^1 tr([\mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k+t\Delta\mathbf{U}) - \mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k)]\Delta\mathbf{U})dt \\ \leq& \|\Delta\mathbf{U}\|_{S^\infty}\| \int_0^1 \mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k+t\Delta\mathbf{U}) - \mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k)\|_{S^1} dt \end{aligned} \tag{14}$$

We focus on the integration part and like first part we use first-order Taylor expansion approximation.

$$\begin{aligned}\int_0^1 \mathbf{V}_{p,:}\triangledown_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k + t\Delta\mathbf{U}) - \mathbf{V}_{p,:}\triangledown_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k) \\ \simeq \int_0^1 \frac{d}{d_t}\mathbf{V}_{p,:}\triangledown_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k + t\Delta\mathbf{U})|_{t=0} t d_t \\ = \frac{1}{2}\frac{d}{d_t}\mathbf{V}_{p,:}\triangledown_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k + t\Delta\mathbf{U})|_{t=0}\end{aligned} \tag{15}$$

From RNN's hidden states iterative equation $\mathbf{s}_t = \sigma(\mathbf{W}\mathbf{s}_{t-1} + \mathbf{U}\mathbf{x}_t)$, we can calculate $\mathbf{U}$'s first order gradient: $\triangledown_{\mathbf{U}}\mathbf{V}_p\mathbf{s}_t = \mathbf{\Lambda}'_t\mathbf{V}_p^T\mathbf{x}_t + \mathbf{Q}_{t-1}\tilde{\mathbf{W}}^T\tilde{\mathbf{\Lambda}}'_t\tilde{\mathbf{V}}_p^T$ where $\mathbf{\Lambda}'_t$ is a diagonal matrix whose diagonal vector is $\sigma'(\mathbf{W}\mathbf{s}_{t-1} + \mathbf{U}\mathbf{x}_t)$. Subsequently, we can obtain the following equation where we ignore the second order part $\frac{d_{\mathbf{Q}_{t-1}}}{d_t} = tr(\frac{\partial^2\mathbf{s}_{t-1}}{\partial\mathbf{U}^2}\Delta\mathbf{U})$.

$$\frac{d}{d_t}\triangledown_{\mathbf{U}}\mathbf{V}_p\mathbf{s}_t = \frac{d}{d_t}\mathbf{\Lambda}'_t\mathbf{V}_p^T\mathbf{x}_t + \mathbf{Q}_{t-1}\tilde{\mathbf{W}}^T\frac{d}{d_t}\tilde{\mathbf{\Lambda}}'_t\tilde{\mathbf{V}}_p^T \tag{16}$$

Then we can obtain the upper bound containing $\|\mathbf{U}\|_{S^\infty}$ of the equation above using equation 8, 9, and 11.

$$\begin{aligned}\|\frac{d}{d_t}\mathbf{\Lambda}'_t\mathbf{V}_p^T\mathbf{x}_t\|_{S^1} \leq \|\frac{d}{d_t}\lambda'_t \odot \mathbf{V}_p^T\|_2\|\mathbf{x}_t\|_2 \leq \|\frac{d}{d_t}\lambda'_t\|_\infty\|\mathbf{V}_p\|_2\|\mathbf{x}_t\|_2 \\ \leq max_q(\lambda'^{-1}_{t_q}\lambda''_{t_q}\|\mathbf{Q}_{t-1_{:,:,q}}\|_{S^1})\|\Delta\mathbf{U}\|_{S^\infty}\|\mathbf{V}_p\|_2\|\mathbf{x}_t\|_2\end{aligned} \tag{17}$$

$$\begin{aligned}\|\mathbf{Q}_{t-1}\tilde{\mathbf{W}}^T\frac{d}{d_t}\tilde{\mathbf{\Lambda}}'_t\tilde{\mathbf{V}}_p^T\|_{S^1} \leq \|\|\mathbf{Q}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\mathbf{W}^T\frac{d}{d_t}\lambda'_t \odot \mathbf{V}_p^T\|_2 \\ \leq \|\|\mathbf{Q}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\mathbf{W}\|_{S^\infty}\|\frac{d}{d_t}\lambda'_t \odot \mathbf{V}_p^T\|_2 \leq \|\|\mathbf{Q}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\mathbf{W}\|_{S^\infty}\|\frac{d}{d_t}\lambda'_t\|_\infty\|\mathbf{V}_p\|_2 \\ \leq \|\|\mathbf{Q}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\mathbf{W}\|_{S^\infty}\max_q(\lambda'^{-1}_{t_q}\lambda''_{t_q}\|\mathbf{Q}_{t-1_{:,:,q}}\|_{S^1})\|\Delta\mathbf{U}\|_{S^\infty}\|\mathbf{V}_p\|_2\end{aligned} \tag{18}$$

The last less than sign of both formulas and the first less than sign of the second one use the following two lemmas respectively.

Lemma 3 $\|\frac{d}{d_t}\lambda'_t\|_\infty \leq \max_q(\lambda'^{-1}_{t_q}\lambda''_{t_q}\|\mathbf{Q}_{t_{:,:,q}}\|_{S^1})$

We notice that $\lambda'_t = \sigma'(\mathbf{W}\mathbf{s}_{t-1} + \mathbf{U}\mathbf{x}_t)$ and $\mathbf{s}_t = \sigma(\mathbf{W}\mathbf{s}_{t-1} + \mathbf{U}\mathbf{x}_t)$, so according to the chain rule of differential calculus, we have the following equation.

$$\frac{\frac{\partial\mathbf{s}_t}{\partial\mathbf{U}}}{\lambda'_t} = \frac{\frac{\partial\lambda'_t}{\partial\mathbf{U}}}{\lambda''_t} \rightarrow \frac{\partial\lambda'_t}{\partial\mathbf{U}} = \mathbf{Q}_t\tilde{\mathbf{\Lambda}}'^{-1}_t\tilde{\mathbf{\Lambda}}''_t \tag{19}$$

So we can derive the upper bound of p-th element of $\frac{d}{d_t}\lambda'_t$ and this is consistent with **Lemma 3**. #

$$[\frac{d}{d_t}\lambda_t]_p = tr[\frac{\partial\lambda'_t}{\partial\mathbf{W}_{:,:,p}}\Delta\mathbf{W}] \leq \|\mathbf{P}_{t-1_{:,:,p}}\lambda'^{-1}_p\lambda''_p\|_{S^1}\|\Delta\mathbf{W}\|_{S^\infty} = \lambda'^{-1}_p\lambda''_p\|\mathbf{P}_{t-1_{:,:,p}}\|_{S^1}\|\Delta\mathbf{W}\|_{S^\infty} \tag{20}$$

Lemma 4 For a 3-d tensor $\mathbf{T}$ and a vector $\mathbf{v}$, we have $\|\mathbf{T}\tilde{\mathbf{v}}\|_{S^1} \leq \|\|\mathbf{T}_{:,:,*}\|_{S^1}\|_2\|\mathbf{v}\|_2$

This lemma can be simply proved using the triangle property of Schatten-1 and Euclidean norm.

$$\|\mathbf{T}\tilde{\mathbf{v}}\|_{S^1} = \|\sum_p \mathbf{v}_p\mathbf{T}_{:,:,p}\|_{S^1} \leq \sum_p \mathbf{v}_p\|\mathbf{T}_{:,:,p}\|_{S^1} \leq \|\|\mathbf{T}_{:,:,*}\|_{S^1}\|_2\|\mathbf{v}\|_2\# \tag{21}$$

Integrate formula 13,14,15,17 and 18, we can obtain the upper bound of Lipschitz constant in Schatten-∞ norm for input matrix $\mathbf{U}$.

$$\begin{aligned} L_{\mathbf{U}} = \frac{1}{2}\max_q \|\mathbf{Q}_t\tilde{\mathbf{V}}_q\|_{S^1}^2 + \max_q \|\mathbf{V}_q\|_2\|\mathbf{x}_t\|_2 max_p(\lambda_p'^{-1}\lambda_p''\|\mathbf{Q}_{t-1_{:,:,p}}\|_{S^1}) \\ + \max_q \|\mathbf{V}_q\|_2\|\mathbf{W}\|_{S^\infty}\|\|\mathbf{Q}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\max_p(\lambda_p'^{-1}\lambda_p''\|\mathbf{Q}_{t-1_{:,:,p}}\|_{S^1}) \end{aligned} \tag{22}$$

B.2 Lipschitz Constant of the Recurrent Matrix

Like the input matrix, we need to upper bound $\|\mathbf{V}\mathbf{s}_t(\mathbf{W}^k+\Delta\mathbf{W}) - \mathbf{V}\mathbf{s}_t(\mathbf{W}^k)\|_\infty$ and $\|\mathbf{V}\mathbf{s}_t(\mathbf{W}^k + \Delta\mathbf{W}) - \mathbf{V}\mathbf{s}_t(\mathbf{W}^k) - \langle\mathbf{V}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k), \Delta\mathbf{W}\rangle\|_\infty$.

Like formula 13, we can use first order approximation to obtain the upper bound of the first part.

$$\|\mathbf{V}\mathbf{s}_t(\mathbf{W}^k + \Delta\mathbf{W}) - \mathbf{V}\mathbf{s}_t(\mathbf{W}^k)\|_\infty \leq \max_p \|\mathbf{P}_t\tilde{\mathbf{V}}_p\|_{S^1}\|\Delta\mathbf{W}\|_{S^\infty} \tag{23}$$

For the second part, like analysis for matrix $\mathbf{U}$, we have the following formulas.

$$\begin{aligned} &[\mathbf{V}\mathbf{s}_t(\mathbf{W}^k + \Delta\mathbf{W}) - \mathbf{V}\mathbf{s}_t(\mathbf{W}^k) - \langle\mathbf{V}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k), \Delta\mathbf{W}\rangle]_p \\ &= \int_0^1 tr([\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k + t\Delta\mathbf{W}) - \mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k)]\Delta\mathbf{W})dt \\ &\leq \|\Delta\mathbf{W}\|_{S^\infty}\| \int_0^1 \mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k + t\Delta\mathbf{W}) - \mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k)\|_{S^1} dt \end{aligned} \tag{24}$$

$$\begin{aligned} &\int_0^1 \mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k + t\Delta\mathbf{W}) - \mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k)dt \\ &\simeq \int_0^1 \frac{d}{d_t}\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k + t\Delta\mathbf{W})|_{t=0}td_t \\ &= \frac{1}{2}\frac{d}{d_t}\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k + t\Delta\mathbf{W})|_{t=0} \end{aligned} \tag{25}$$

Unlike scenarios for $\mathbf{U}$, $\frac{d}{d_t}\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k + t\Delta\mathbf{W})|_{t=0}$ is complicated and contains four parts.

$$\triangledown_{\mathbf{W}}(\mathbf{V}_p\mathbf{s}_t) = (\bar{\mathbf{S}}_{t-1} + \mathbf{P}_{t-1}\tilde{\mathbf{W}}^T)\tilde{\mathbf{\Lambda}'}_t\tilde{\mathbf{V}}_p^T = \mathbf{\Lambda'}_t\mathbf{V}_p^T\mathbf{s}_{t-1} + \mathbf{P}_{t-1}\tilde{\mathbf{W}}^T\tilde{\mathbf{\Lambda}'}_t\tilde{\mathbf{V}}_p^T \tag{26}$$

$$\frac{d}{d_t}\triangledown_{\mathbf{W}}(\mathbf{V}_p\mathbf{s}_t) = \frac{d}{d_t}\mathbf{\Lambda'}_t\mathbf{V}_p^T\mathbf{s}_{t-1} + \mathbf{\Lambda'}_t\mathbf{V}_p^T\frac{d}{d_t}\mathbf{s}_{t-1} + \mathbf{P}_{t-1}\frac{d}{d_t}\tilde{\mathbf{W}}^T\tilde{\mathbf{\Lambda}'}_t\tilde{\mathbf{V}}_p^T + \mathbf{P}_{t-1}\tilde{\mathbf{W}}^T\frac{d}{d_t}\tilde{\mathbf{\Lambda}'}_t\tilde{\mathbf{V}}_p^T \tag{27}$$

Similarly, we have to find the upper bound of each part using $\|\Delta\mathbf{W}\|_{S^\infty}$.

$$\begin{aligned} &\|\frac{d}{d_t}\mathbf{\Lambda'}_t\mathbf{V}_p^T\mathbf{s}_{t-1}\|_{S^1} = \|(\frac{d}{d_t}\lambda_t' \odot \mathbf{V}_p^T)\mathbf{s}_{t-1}\|_{S^1} \leq \|\frac{d}{d_t}\lambda_t' \odot \mathbf{V}_p^T\|_2\|\mathbf{s}_{t-1}\|_2 \\ &\leq \|\frac{d}{d_t}\lambda_t'\|_\infty\|\mathbf{V}_p\|_2\|\mathbf{s}_{t-1}\|_2 = \max_q(\lambda_{t_q}'^{-1}\lambda_{t_q}''\|\mathbf{P}_{t-1_{:,:,q}}\|_{S^1})\|\Delta\mathbf{W}\|_{S^\infty}\|\mathbf{V}_p\|_2\|\mathbf{s}_{t-1}\|_2 \end{aligned} \tag{28}$$

$$\begin{aligned} \|\mathbf{\Lambda}_t'\mathbf{V}_{p,:}^T(\frac{d}{d_t}\mathbf{s}_{t-1}^T)\|_{S^1} &\leq \|\lambda_t' \odot \mathbf{V}_p^T\|_2\|\frac{d}{d_t}\mathbf{s}_{t-1}\|_2 \leq \|\lambda_t'\|_\infty\|\mathbf{V}_p\|_2\|tr[\frac{\partial\mathbf{s}_{t-1}}{\partial\mathbf{W}}\Delta\mathbf{W}]\|_2 \\ &= \|\lambda_t'\|_\infty\|\mathbf{V}_p\|_2\|\Delta\mathbf{W}\|_{S^\infty}\|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2 \end{aligned} \tag{29}$$

$$
\begin{aligned}
\|\mathbf{P}_{t-1}\frac{d}{d_t}\tilde{\mathbf{W}}^T\tilde{\mathbf{\Lambda}}'_t\tilde{\mathbf{V}}_p^T\|_{S^1} &\leq \|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\frac{d}{d_t}\mathbf{W}^T\mathbf{\Lambda}'_t\mathbf{V}_p^T\|_2 \\
= \|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\Delta\mathbf{W}^T\mathbf{\Lambda}'_t\mathbf{V}_p^T\| &\leq \|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\Delta\mathbf{W}\|_{S^\infty}\|\lambda'_t\odot\mathbf{V}_p^T\|_2 \\
&\leq \|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\Delta\mathbf{W}\|_{S^\infty}\|\lambda'_t\|_\infty\|\mathbf{V}_p^T\|_2
\end{aligned} \tag{30}
$$

$$
\begin{aligned}
\|\mathbf{P}_{t-1}\tilde{\mathbf{W}}^T\frac{d}{d_t}\tilde{\mathbf{\Lambda}}'_t\tilde{\mathbf{V}}_p^T\|_{S^1} &\leq \|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\mathbf{W}^T\frac{d}{d_t}\mathbf{\Lambda}'_t\mathbf{V}_p^T\|_2 \\
\leq \|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\mathbf{W}\|_{S^\infty}\|\frac{d}{d_t}\lambda'_t\odot\mathbf{V}_p^T\| &\leq \|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\mathbf{W}\|_{S^\infty}\|\frac{d}{d_t}\lambda'_t\|_\infty\|\mathbf{V}_p\|_2 \\
&\leq \|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2\|\mathbf{W}\|_{S^\infty}\max_q(\lambda'^{-1}_{t_q}\lambda''_{t_q}\|\mathbf{P}_{t-1_{:,:,q}}\|)\|\Delta\mathbf{W}\|_{S^\infty}\|\mathbf{V}_p\|_2
\end{aligned} \tag{31}
$$

Combined formulas above, we can obtain the Lipschitz constant in Schatten-∞ norm for recurrent matrix $\mathbf{W}$.

$$
\begin{aligned}
L_{\mathbf{W}} = \frac{1}{2}(\max_q\|\mathbf{P}_t\tilde{\mathbf{V}}_q\|_{S^1})^2 + \max_q\|\mathbf{V}_q\|_2\|\mathbf{s}_{t-1}\|_2\max_p(\lambda'^{-1}_p\lambda''_p\|\mathbf{P}_{t-1_{:,:,p}}\|_{S^1}) \\
+2\max_q\|\mathbf{V}_q\|_2\|\lambda'_t\|_\infty\|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2 + \max_q\|\mathbf{V}_q\|_2\|\mathbf{W}\|_{S^\infty}\|\|\mathbf{P}_{t-1_{:,:,*}}\|_{S^1}\|_2\max_p(\lambda'^{-1}_p\lambda''_p\|\mathbf{P}_{t-1_{:,:,p}}\|)
\end{aligned} \tag{32}
$$

C References

[1]. Carlson, David E., Volkan Cevher, and Lawrence Carin. "Stochastic Spectral Descent for Restricted Boltzmann Machines." AISTATS. 2015.

[2]. Bernstein, Dennis S. *Matrix mathematics: theory, facts, and formulas.* Princeton University Press, 2009.

[3]. Carlson, David E., et al. *"Preconditioned spectral descent for deep learning."* Advances in Neural Information Processing Systems. 2015.