Orthogonal Recurrent Neural Networks with Scaled Cayley Transform

Supplemental Material: Proof of Theorem 3.2

For completeness, we restate and prove Theorem 3.2.

Theorem 3.2 *Let $L = L(W) : \mathbb{R}^{n\times n} \to \mathbb{R}$ be some differentiable loss function for an RNN with the recurrent weight matrix W. Let $W = W(A) := (I+A)^{-1}(I-A)D$ where $A \in \mathbb{R}^{n\times n}$ is skew-symmetric and $D \in \mathbb{R}^{n\times n}$ is a fixed diagonal matrix consisting of -1 and 1 entries. Then the gradient of $L = L(W(A))$ with respect to A is*

$$\frac{\partial L}{\partial A} = V^T - V \tag{1}$$

where $V := (I+A)^{-T}\frac{\partial L}{\partial W}\left(D + W^T\right)$, $\frac{\partial L}{\partial A} = \left[\frac{\partial L}{\partial A_{i,j}}\right] \in \mathbb{R}^{n\times n}$, and $\frac{\partial L}{\partial W} = \left[\frac{\partial L}{\partial W_{i,j}}\right] \in \mathbb{R}^{n\times n}$

Proof: Let $Z := (I+A)^{-1}(I-A)$. We consider the (i,j) entry of $\frac{\partial L}{\partial A}$. Taking the derivative with respect to $A_{i,j}$ where $i \neq j$ we obtain:

$$\begin{aligned}\frac{\partial L}{\partial A_{i,j}} &= \sum_{k,l=1}^{n}\frac{\partial L}{\partial W_{k,l}}\frac{\partial W_{k,l}}{\partial A_{i,j}} = \sum_{k,l=1}^{n}\frac{\partial L}{\partial W_{k,l}}D_{l,l}\frac{\partial Z_{k,l}}{\partial A_{i,j}}\\ &= \operatorname{tr}\left[\left(\frac{\partial L}{\partial W}D\right)^T\frac{\partial Z}{\partial A_{i,j}}\right]\end{aligned}$$

Using the identity $(I+A)Z = I - A$ and taking the derivative with respect to $A_{i,j}$ to both sides we obtain:

$$\frac{\partial Z}{\partial A_{i,j}} + \frac{\partial A}{\partial A_{i,j}}Z + A\frac{\partial Z}{\partial A_{i,j}} = -\frac{\partial A}{\partial A_{i,j}}$$

and rearranging we get:

$$\frac{\partial Z}{\partial A_{i,j}} = -(I+A)^{-1}\left(\frac{\partial A}{\partial A_{i,j}} + \frac{\partial A}{\partial A_{i,j}}Z\right)$$

Let $E_{i,j}$ denote the matrix whose (i,j) entry is 1 with all others being 0. Since A is skew-symmetric, we have $\frac{\partial A}{\partial A_{i,j}} = E_{i,j} - E_{j,i}$. Combining everything, we have:

$$\begin{aligned}\frac{\partial L}{\partial A_{i,j}} &= -\operatorname{tr}\left[\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}\left(E_{i,j} - E_{j,i} + E_{i,j}Z - E_{j,i}Z\right)\right]\\ &= -\operatorname{tr}\left[\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}E_{i,j}\right]\\ &\quad + \operatorname{tr}\left[\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}E_{j,i}\right]\\ &\quad - \operatorname{tr}\left[\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}E_{i,j}Z\right]\\ &\quad + \operatorname{tr}\left[\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}E_{j,i}Z\right]\\ &= -\left[\left(\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}\right)^T\right]_{i,j}\\ &\quad + \left[\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}\right]_{i,j}\\ &\quad - \left[\left(\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}\right)^T Z^T\right]_{i,j}\\ &\quad + \left[Z\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}\right]_{i,j}\\ &= \left[(I+Z)\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}\right]_{i,j}\\ &\quad - \left[\left(\left(\frac{\partial L}{\partial W}D\right)^T(I+A)^{-1}\right)^T\left(I+Z^T\right)\right]_{i,j}\\ &= \left[(D+W)\left(\frac{\partial L}{\partial W}\right)^T(I+A)^{-1}\right]_{i,j}\\ &\quad - \left[(I+A)^{-T}\frac{\partial L}{\partial W}\left(D+W^T\right)\right]_{i,j}\end{aligned}$$

Using the above formulation, $\frac{\partial L}{\partial A_{j,j}} = 0$ and $\frac{\partial L}{\partial A_{i,j}} = -\frac{\partial L}{\partial A_{j,i}}$ so that $\frac{\partial L}{\partial A}$ is a skew-symmetric matrix. Finally, by the definition of V we get the desired result. ■