# Supplementary Materials

## 1. Activation Function

**sigmoid function:**

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\sigma'(x) = \sigma(x) \cdot [1 - \sigma(x)]$$

**hyperbolic function:**

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$\tanh'(x) = 1 - \tanh^{2}(x)$$

**rectified linear unit(ReLU):**

$$f(x) = \max(0, x)$$

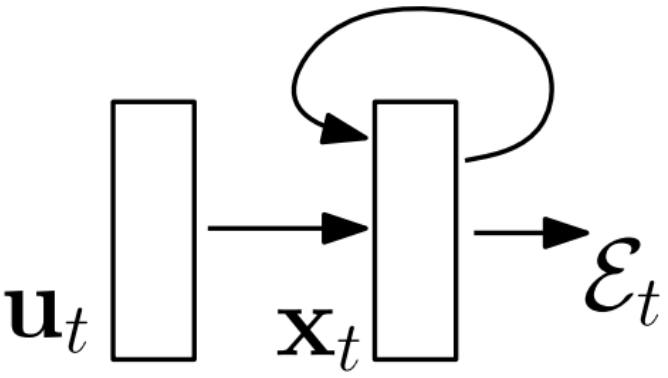
**softmax function:**

$$\mathbf{y} = \text{softmax}(\mathbf{x})$$

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

$$\frac{\partial y_i}{\partial x_j} = \begin{cases} -y_i \cdot y_j, & i \neq j \\ y_i \cdot (1 - y_i), & i = j \end{cases}$$

## 2. Vanishing Gradient in RNN [1]



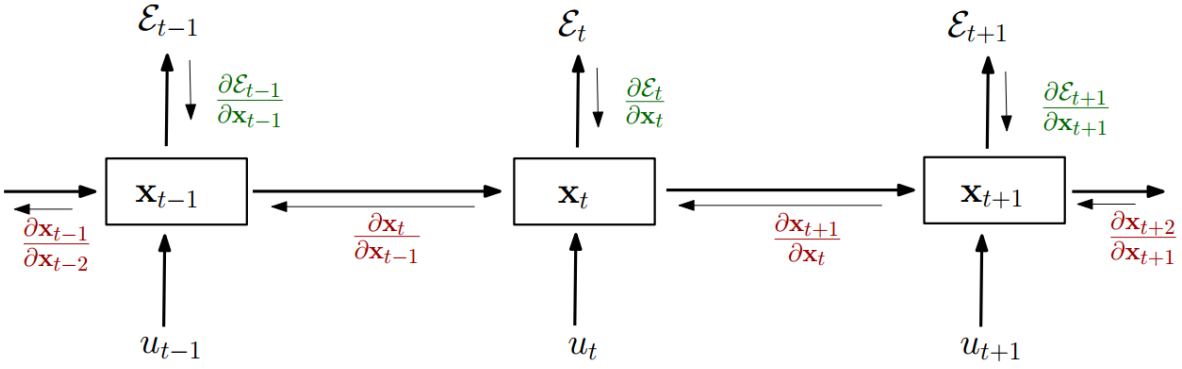

**hidden state:**

$$\mathbf{x}_t = \mathbf{W}_{rec}\sigma(\mathbf{x_{t-1}}) + \mathbf{W}_{in}\mathbf{u}_t + \mathbf{b}$$

**cost:**

$$\mathcal{E} = \sum_{1 \le t \le T} \mathcal{E}_t = \sum_{1 \le t \le T} \mathcal{L}(\mathbf{x}_t)$$

**unrolling RNN:**




**gradients:**

$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left( \frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right)$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T \mathrm{diag}(\sigma'(\mathbf{x}_{i-1}))$$

**proof**:

it is sufficient for $\lambda_1 < \frac{1}{\gamma}$, where $\lambda_1$ is the largest singular value of $\mathbf{W}_{rec}$ and $\left\|\text{diag}(\sigma'(\mathbf{x}_k))\right\| \leq \gamma \in \mathcal{R}$, for the vanishing gradient problem to occur.

$$\forall k, \left\|\frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k}\right\| \leq \left\|\mathbf{W}_{rec}^T\right\| \left\|\text{diag}(\sigma'(\mathbf{x}_k))\right\| < \frac{1}{\gamma}\gamma < 1$$

let $\eta \in \mathcal{R}$ be such that $\forall k, \left\|\frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k}\right\| \leq \eta < 1$.

$$\left\|\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t}\left(\prod_{i=k}^{t-1}\frac{\partial \mathbf{x}_{i+1}}{\partial \mathbf{x}_i}\right)\right\| \leq \eta^{t-k}\left\|\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t}\right\|$$

**deal with the exploding and vanishing gradient**:

- $L1$ or $L2$ penalty
- LSTM
- clipping gradient

**gradient flow in LSTM**:

$$\frac{\partial \mathbf{c}_t}{\partial \mathbf{c}_k} = \frac{\partial \mathbf{c}_t}{\partial \mathbf{c}_{t-1}} \cdots \frac{\partial \mathbf{c}_{k+1}}{\partial \mathbf{c}_k} = \text{diag}(\mathbf{f}_t) \cdots \text{diag}(\mathbf{f}_k) = \text{diag}(\mathbf{f}_t \odot \cdots \odot \mathbf{f}_k)$$

# Reference


1. **Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." Proceedings of The 30th International Conference on Machine Learning. 2013.** http://www.jmlr.org/proceedings/papers/v28/pascanu13.pdf