On penalty terms

Giulio Galvan

28 settembre 2015

We have seen that:

$$\frac{\partial g}{\partial W^{rec}} = \sum_{t=1}^{T} \frac{\partial g_t}{\partial \boldsymbol{a}^t} \cdot \sum_{k=1}^{t} \frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k} \cdot \frac{\partial \boldsymbol{a}^k}{\partial W^{rec}}, \tag{1}$$

where

$$\frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k} = \prod_{i=t-1}^{k} diag(\sigma'(\boldsymbol{a}^i)) \cdot W^{rec}. \tag{2}$$

tends to vanish.

The idea is to develop a penalty term which express a preference for solutions where the components $\frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k}$ are far from zero, hence, learning a model which exhibit long memory.

A first attempt could be:

$$\Gamma \triangleq \sum_{t=1}^{T} \sum_{k=1}^{t} \frac{1}{\left\| \frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k} \right\|^2}. \tag{3}$$

Γ treats all temporal steps equally: we could assign more importance to distant temporal steps, which are the most critical, modifying $Gamma$ as follows:

$$\Gamma \triangleq \sum_{t=1}^{T} \sum_{k=1}^{t} \frac{\sigma(t-k)}{\left\| \frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k} \right\|^2}, \tag{4}$$

where $\sigma(\cdot)$ assign different weights depending on the temporal distance $t-k$, for example $\sigma(h) = exp\{h\}$